An Investigation into Deviant Morphology: Issues in the Implementation of a Deep Grammar for Indonesian

A thesis submitted
for the degree of
Doctor of Philosophy
of
The Australian National University
Canberra, Australia

Meladel Mistica

June 2013

Declaration

This is to certify that:

(i) the thesis comprises only my original work towards the PhD except where indicated in the Preface;

(ii) due acknowledgement has been made in the text to all other material used;

(iii) the thesis is fewer than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Signed: _____ Date: _____

Thesis supervisor(s)
I Wayan Arka
Timothy Baldwin

Author
Meladel Mistica

An Investigation into Deviant Morphology: Issues in the Implementation of a Deep Grammar for Indonesian

Abstract

This thesis investigates deviant morphology in Indonesian for the implementation of a deep grammar. In particular we focus on the implementation of the verbal suffix *-kan*. This suffix has been described as having many functions, which alter the kinds of arguments and the number of arguments the verb takes (Dardjowidjojo 1971; Chung 1976; Arka 1993; Vamarasi 1999; Kroeger 2007; Son and Cole 2008).

Deep grammars or precision grammars (Butt *et al.* 1999a; Butt *et al.* 2003; Bender *et al.* 2011) have been shown to be useful for natural language processing (NLP) tasks, such as machine translation and generation (Oepen *et al.* 2004; Cahill and Riester 2009; Graham 2011), and information extraction (MacKinlay *et al.* 2012), demonstrating the need for linguistically rich information to aid NLP tasks. Although these linguistically-motivated grammars are invaluable resources to the NLP community, the biggest drawback is the time required for the manual creation and curation of the lexicon. Our work aims to expedite this process by applying methods to assign syntactic information to *kan*-affixed verbs automatically. The method we employ exploits the hypothesis that semantic similarity is tightly connected with syntactic behaviour (Levin 1993).

Our endeavour in automatically acquiring verbal information for an Indonesian deep grammar poses a number of lingustic challenges. First of all Indonesian verbs exhibit *voice* marking that is characteristic of the subgrouping of its language family. In order to be able to characterise verbal behaviour in Indonesian, we first need to devise a detailed analysis of voice for implementation. Another challenge we face is the claim that all open class words in Indonesian, at least as it is spoken in some varieties (Gil 1994; Gil 2010), cannot linguistically be analysed as being distinct from each other. That is, there is no distiction between nouns, verbs or adjectives in Indonesian, and all word from the open class categories should be analysed uniformly. This poses difficulties in implementing a grammar in a linguistically motivated way, as well discovering syntactic behaviour of verbs, if verbs cannot be distinguished from nouns. As part of our investigation we conduct experiments to verify the need to employ word class categories, and we find that indeed these are linguistically motivated labels in Indonesian.

Through our investigation into deviant morphological behaviour, we gain a better characterisation of the morphosyntactic effects of *-kan*, and we discover that, although Indonesian has been labelled as a language with no open word class distinctions, word classes can be established as being linguistically-motivated.

Contents

List of Figures

List of Tables

Abbreviations and Glossing Conventions

Throughout the thesis we employ the *Conventions for interlinear morpheme-by-morpheme glosses*, also known as the *Leipzig Glossing Rules*, as formulated by the Max Planck Institute for Evolutionary Anthropology.[1] However, we deviate from the Leipzig glossing rules with respect to reduplication. The reduplicate in the surface (Indonesian) word should be introduced with a *tilde* symbol '~', but we represent this with a *hyphen* '-' to reflect the Indonesian orthographic conventions for full reduplication, for example *orang-orang* "person~person" or *orang-orang* "people", and not *orang~orang* "orang~PL". The abbreviations we employ are listed below:

1, 2, 3	1st, 2nd, 3rd person
A	agent
CLF	classifier
DEF	definite
EMPH	emphatic
EXCL	exclusive
HON	honorific
INCL	inclusive
PASS	passive
REFL	reflexive
REL	relative

In addition, we use the glossing conventions of other authors when citing works for other languages. For Tagalog, Foley (2008) uses the following conventions:

VC	voice marker
CORE	core argument
IRR	irrealis

Teng (2008), Arka (2008), and Jukes (2012) employ the following for Puyuma, Balinese, and Makassarese, respectively:

APPL	applicative
DF	definite
ID	indefinite
IND	indicative
ITR	intransitive
POSS	possessive
PREP	preposition
PRS	present
RED	reduplicate

Also, Teng (2008) delimits so-called voice marking in Puyuma within angle brackets '⟨ ⟩', and Foley (2008) delimits partial reduplicated forms with a hyphen '-'.

[1] Details can be found here: `http://www.eva.mpg.de/lingua/resources/glossing-rules.php`

Citations to Previously Published Work

Large portions of Chapter 5 have appeared in the following paper:

> Mistica *et al.* (2011) Word Classes in Indonesian: A Linguistics Reality or a Convenient Fallacy in Natural Language Processing?, In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, NTU, Singapore pp. 293–302.

Portions of Section 2.2 have appeared in the following paper:

> Mistica *et al.* (2009) Double Double, Morphology and Trouble: Looking into Reduplication in Indonesian, In *Proceedings of the 2009 Australasian Language Technology Workshop (ALTW 2009)*, Sydney Australia, pp. 44–52.

Large portions of Chapter 6 have appeared in the following paper:

> Mistica *et al.* (2013) Unsupervised Word Class Induction for Under-resourced Languages: A Case Study on Indonesian, In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, Nagoya, Japan, pp. 685–691.

Acknowledgments

I am extremely grateful to my supervisors Wayan Arka and Tim Baldwin for their invaluable advice and support throughout my candidature. You are superstars. I'd also like to thank the rest of my panel Avery Andrews and Jane Simpson who were also a great support, and equally superstars.

Thanks to the folks at ANU and Melbourne Uni. I have been lucky enough to be a part of two great departments with such lovely people.

I am also very grateful to have had the opportunity to attend ParGram meetings and to sit and learn from the greats. Another instance where I've been lucky enough to be amongst a swell bunch.

Thank you very much to the three examiners who generously took the time to provide insightful comments and welcome suggestions.

This project was made possible by the ARC Grant DP0877595, which allowed me to attend conferences and meetings, and also provided a salary.

Finally thanks to my family and friends, both here and afar. Thanks for the visits, coffees, Roo spag bol, "Meladel" dinners, and cake! Thank you for your kind and supportive words, and for your patience (and chocolate). And a special big thanks to my mama. Eta Txori-buru, eskerrik asko!

For my dad.

Part I

Setting the Scene

Chapter 1

Introduction

In Martin Kay's speech upon accepting his Lifetime Achievement award for *The Association for Computational Linguistics* in 2005, he defined the research areas of **Computational Linguistics** and **Natural Language Processing** as such:

> Computational linguistics is not natural language processing. Computational linguistics is trying to do what linguists do in a computational manner, not trying to process texts, by whatever methods, for practical purposes. Natural Language Processing, on the other hand, is motivated by engineering concerns. I suspect that nobody would care about building probabilistic models of language unless it was thought that they would serve some practical end. There is nothing unworthy in such an enterprise.

However, Natural Language Processing (NLP) and Computational Linguistics (CL) need not be mutually exclusive, and indeed since the time Kay delivered this ACL Lifetime Achievement Award speech there has been a great effort in closing the gap between the two disciplines.

This thesis explores how computational linguistics, and linguistic analysis, can contribute to the goals of natural language processing, specifically in the creation of linguistically motivated resources, and in the design of features for the application of machine learning tasks. In turn, we also investigate how Deep Lexical Acquisition (DLA) and the application of these NLP methods can inform linguistic inquiry, by exploiting the large amounts of non-elicited evidence of usage by online content creators.

Specifically, this work aims to contribute to a *deep grammar* for Indonesian. The focus of this work lies within the sublexical domain of Indonesian. We investigate aspects of verbal morphology, in particular the suffix *-kan*, which can alter the arguments a verb requires (*predicate-argument-changing morphology*). While typically valence-changing, this suffix may not always result in an increase in valency, as we discuss in the later chapters. This morphosyntactic topic has been studied by Austronesianists and linguists in depth for at least the last four decades (Dardjowidjojo 1971;

Chung 1976; Muhadjir 1981; Arka 1993; Kroeger 2007; Arka *et al.* 2009), making it ideal for computational linguistic investigation.

Morphological processes can increase or decrease the number of direct arguments a verb takes. We aim to better understand the mechanisms that license these changes from a linguistic perspective, so that we may be able to encode them in a systematic and meaningful way in the deep grammar. We consider the English passive as one kind of argument-changing process.

(1.1) *Sue* coaxed *Mary* (into skydiving).

(1.2) *Mary* **was** coax**ed** (into skydiving).

Examples (1.1) and (1.2) show a decrease in the number of direct arguments that are expressed.[1] In Example (1.1), there are two participants that are both overtly expressed as arguments. However, in Example (1.2) we see that with changes to the form of the verb, only one argument to the verb needs to be expressed. There is no dedicated morpheme that triggers the passive, but it is constructed with the auxiliary verb *be* and the suffix *-ed* on the verb.

Unlike the passive construction, there is no overt morphological marking that signals the changes the number of arguments a verb can take in English.

(1.3) *Sue* bought *a skydiving voucher* (for *Mary*).

(1.4) *Sue* bought *Mary a skydiving voucher.*

The relationship between the verbs in Examples (1.3) and (1.4) and whether they should be represented as a single lexical entry or two different ones has been called to question (Levin 1993). There is no overt morphological difference between the verbs shown in these examples, however in the latter example, there are three direct arguments, and only two in the former. This phenomenon is referred to as a *diathesis alternation*, and Example (1.4) is the *benefactive alternation*. These are alternative realisations of arguments for a verb that is sometimes accompanied with a slight shift in meaning (Levin 1993). The benefactive in Indonesian is triggered by the suffix *-kan*, and although not a complete description of the suffix, it is commonly characterised as having two major functions: (1) benefactive applicativisation; and (2) causativisation (Arka 1993; Kroeger 2007). The process of applicativising adds an extra non-subject argument, as shown in Example (1.5).

(1.5) (Kaswanti 1997 via Kroeger 2007)

a. *John membeli buku itu untuk Mary.*
J AV-buy book this for M
"John bought a book for Mary."

[1]The parentheses indicate optionality.

b. *John membelikan Mary buku itu.*
J AV-buy-KAN M book this
"John bought Mary a book."

c. **John membeli Mary buku itu.*
J AV-buy M book this
"John bought Mary a book."

The following is an example of the suffix *-kan* triggering a causative construction, with the 'added' argument taking the role of the 'causer':

(1.6) a. *Orang-orang mengungsi.*
person~person AV-take.refuge
"The people took refuge."

b. *PBB mengungsikan orang-orang.*
U.N. AV-take.refuge-KAN person-person
"The U.N. evacuated the people."

We investigate the linguistic mechanisms that license these syntactic changes within a *lexicalist framework*, namely Lexical Functional Grammar, which we introduce in Chapter 2.

By delving into the linguistic facts of these valence-changing phenomena, we devise an implementation for our Indonesian deep grammar. Deep grammars or precision grammars (Butt *et al.* 1999b; Butt and King 2003; Bender *et al.* 2011) are a resource utilised in syntactic parsing to obtain an informed representation of language. Parsing provides syntactic analyses that determines how sentential elements are related to each other for a range of linguistic phenomena (MacKinlay 2012). They have been shown to assist in certain natural language processing (NLP) tasks, such as machine translation and generation (Graham 2011; Cahill and Riester 2009), and information extraction (MacKinlay 2012), demonstrating the need for linguistically rich information to aid such tasks. Although these are precise resources, as has been noted by Kay (2005), there is a time and effort trade-off in producing these; they are labour intensive to produce, and often brittle. This thesis does not address the brittleness of such deep grammars,[2] however, we aim at finding a way to employ stochastic methods used in NLP to expedite the production of the lexicon. We use methods in Deep Lexical Acquisition in this endeavour (Baldwin 2005; Baldwin 2007). Deep Lexical Acquisition encompasses a collection of NLP methods that aim to expand the coverage of a precision grammar (deep grammar) or deep lexical resource.

[2]See Cahill (2004); Forst (2007), who integrate stochastic parsing in symbolic based systems.

1.1 Thesis Structure

The thesis is divided into four main parts: In Part I, we introduce the reader to the area of investigation and explain the background required to understand the contributions made in Part II and III of the thesis, and the resources used in the study. Part II involves the linguistic exploration and characterisation of Indonesian verbal morphology, and aspects of their implementation in the deep grammar. Part III involves the application of DLA to learn stem types in order to map out morpho-syntactic variation imposed by the suffix *-kan*.

The thesis is structured as follows.

Chapter 2 In this background chapter, we introduce *Lexical Functional Grammar* (LFG), which is the formalism we use in implementing the deep grammar.

LFG is a lexically-driven grammar formalism that has multiple distinct, but parallel, levels of representation, which enables the capturing of cross-linguistic variation, as well as similarities. It is the formalism underlying the *language engineering* platform, XLE, we use in our implementation.

Like many Austronesian languages, Indonesian exhibits verbal marking called *voice*, which we introduce in this chapter. Voice marking can be thought of as a thematic coindexing of the grammatical subject, however its spectrum of variation differs from language to language. The study into predicate argument structure (PAS) changing morphology would not be complete without an introduction into the Indonesian *voice* system because it is an integral part of the Indonesian verb. In our schema all clauses have a *voice* feature because it indicates how arguments in the clause are linked to the verb. We then conduct a review of the characterisations of *-kan*, which has been shown to be rather varied; the linguistic analyses that have been presented in the past all deviate from each other. Finally, this background chapter concludes with an outline of the machine learning methods used within DLA, which we employ in the latter part of the thesis.

Chapter 3 describes the resources used in this study, beginning with the grammar engineering platform, and then the off-the-shelf implementations of machine learning algorithms we employ.

Chapter 4 We detail the implementation issues in building a linguistically motivated and linguistically sound computational resource. In particular, we step through some of the issues with respect to implementing voice and the PAS changing affix *-kan*.

We then empirically map out the different behaviour of the affix *-kan* from corpus evidence with the aim to discover types of stems such that when *-kan* is applied, the same morphosyntatic effects apply for each member of that type. Although *-kan* sometimes applicativises, at times causativises, at times seemingly decreases the

number of argument a verb takes, or has no effect at all, all of these variations are not applicable to all stems. Therefore, in order to prevent the overapplication of *-kan* and avoid overgeneration in the lexicon, we map out the possible alternations of various stems when affixed with *-kan*, so that we may be able to generalise the behaviour of this suffix according to the kind of stem it attaches to. In particular, we investigate the syntactic behaviour of 100 verbs and their semantic decomposition, so that we can discover a small number of classes in order to create lexical templates for them, according to their stem types.

Chapter 5 The notion of word classes, such as nouns, verbs and adjectives, is fundamental in both linguistics and computational linguistics. Word classes are the basis for the labels in part-of-speech tagging, and also the building blocks for parsing. In linguistics, they are considered the categories that shape the organisation of the language. In grammar engineering, they are the primitives upon which context-free grammar rules are written, and indeed the categories that we had built our grammar implementation upon. However the notion of word classes in Indonesian has been in question for a number of decades (Gil 1994; Gil 2001; Gil 2010), and in this chapter we undertake a study addressing this very issue. One of the goals we had in this thesis, was to apply lexical acquisition to verbs in Indonesian to help expedite lexicon development, and to mitigate the over-application of *-kan* in the lexicon. But in order to do this, we would have to establish that there is a class of verbs to apply this process to, and therefore we embark on an experiment to see if we can establish these word classes in Indonesian.

The outcome of this investigation has consequences beyond the implementation details of Indonesian, and our endeavour to have our encoding of the grammar reflect linguistic facts. This study also has typological consequences. These word categories may not align across languages: what is expressed as a verb in one language may be expressed as an adjective or noun in another. But one linguistic universality hypothesis that remains despite these variations is that the categories *noun* and *verb* exist in all languages (Croft 2003). In this chapter we describe our experiment that applies an unsupervised data-driven approach to determine whether we can automatically ascertain a noun–verb distinction, and in doing so, shed light on whether Indonesian conforms to Croft's noun–verb universality hypothesis.

Chapter 6 For these lexically rich deep grammar, an important aspect in the development of a deep grammar is the lexicon. Although these linguistically-motivated grammars are invaluable resources for the NLP community, the biggest drawback is the time required for the manual creation and curation of the lexicon. The case study conducted in Chapter 6 aims to expedite this process by automatically assigning syntactic information to stems that make up the verbal elements, on the basis of the predicting of semantic clusters based on distributional similarity. This case

study exploits Levin's (1989) hypothesis that there is a tight connection between the semantics of a verb and its syntactic profile. We test the viability of inferring syntactic information from distributional or contextual semantics, as a proxy for lexical semantics. Our experiments show that, although semantically alike words can be determined using syntactic features, employing semantic methods do not convincingly predict syntactic features.

Chapter 7 In this chapter we suggest ways we can build upon the preliminary work we have conducted for Indonesian within the field of computational linguistics. In particular, we learn from the methods we apply in Chapter 6, and suggest the way forward in further conducting deep lexical acquisition for Indonesian.

1.2 Contributions

In this thesis we add to the increasing knowledge of grammar engineering implementation, particularly for the Austronesian language family. In addition, we update the phrase structure for Indonesian, proposed by Arka and Manning (2008) based on coordination facts by Musgrave (2001). We provide new methods in the study of morpho-syntax, particularly into the much-studied Indonesian suffix *-kan* (Kroeger 2007; Cole and Son 2004; Arka 1993; Chung 1976), even though the changes to the argument structure imposed by *-kan* cannot be predicted using solely semantic similarity or semantic methods. Finally we conduct a study showing the need for word classes in Indonesian, although it is claimed that as it is spoken in some regions (Riau and Jakarta) it exhibits no word classes (Gil 1994; Gil 2001; Gil 2010). Himmelmann (2008) shows that it is possible for a language to have a two-tiered categorisation of word classes – at the syntactic and the sublexical level. In a complementary study to the Yoder (2010) word order study, we show that also at the sublexical domain, Indonesian exhbits word class categories.

More generally, we have learned that, while detailed qualitative analysis is all-important in the field of linguistics, contributions can be made with large scale modelling. And given that language is not always categorical (Keller 2001), stochastic approximations are reliable descriptions of the language, as shown with the dative alternation in English (Bresnan and Nikitina 2008). In this research, we demonstrate that the methods employed in NLP can be used to aid in linguistic investigation. In particular, our work supports the claim of Yoder (2010) who disputes Gil's (1994, 2001) claim that there are no parts of speech in a variety of spoken Indonesian. Furthermore, we show that clustering based derived semantic properties has the potential to predict deep syntactic lexical properties, and with further investigation could be of assistance in semi-automatically constructing a deep lexical resource for a language such as Indonesian, which has limited but growing resources for natural language processing.

Chapter 2

Background

2.1 Introduction

This chapter outlines the background required to understand this thesis. It introduces relevant morphological and syntactic aspects of Indonesian, the language under investigation, in order to follow the implementation details of our contribution to the Indonesian grammar and lexicon resource discussed in Chapter 4. We also show that in terms of NLP, it is not as richly resourced as other languages with the same number of speakers worldwide and presence on the world wide web. In Section 2.2 we introduce the Indonesian language, and in Section 2.3 the unification-based grammar formalism we employ, and the linguistic knowledge required in utilising the grammar engineering tools in Chapter 3.

The two sections following cover methods that are used within this thesis. In particular, Section 2.6 outlines the stochastic and machine learning algorithms employed in DLA to investigate linguistic properties and features of Indonesian in the latter chapters of this thesis, where we devise a method for learning syntactic information via a proxy for lexical semantics. This section also outlines the methods used in the investigation of word classes, which we conduct to ensure the linguistic soundness of the implementation of the grammatical resources described in Chapter 4.

2.2 Indonesian and its Verbal Morphology

This section looks at verbal morphology and we focus on a particular suffix that can affect argument structure, namely *-kan*. In addition, we look at Austronesian voice because this also affects how arguments are realised. We also describe linguistic properties in some detail to assist the reader in better interpreting the glossed examples throughout the thesis.

In addition, we introduce two aspects in which Indonesian morphology seems to generate problematic linguistic analyses. The first is in Section 2.2.4 where we

investigate the various analyses of the suffix *-kan*. In the linguistic community there seems to be no agreed upon treatment of *-kan* with, not only the suffix displaying a huge variety of constructions, but it also seems that many of the analyses of *-kan* deviate from each other. Another aspect where Indonesian morphology gives rise to controversy is in terms of word classes, discussed in Section 2.4. In this section we discuss the claim that in certain varieties of Indonesian there is simply no distinction between open class categories (Gil 1994; Gil 2001; Gil 2005; Gil 2010), which deviates from the claim that all languages minimally distinguish between nouns and verbs (Croft 2003).

2.2.1 About Indonesian

Bahasa Indonesia "Indonesian" has approximately 23 million L1 speakers and given its status as the national language, its spoken by at least 8 times as many throughout Indonesia and migrant populations throughout the world (Gordon 2005). Also depending on how you choose to cut the divide between languages and dialects, it can be said to be spoken by many millions more, with Quinn (2001) describing Indonesian as "the 20th century name for Malay" (Quinn 2001:viii) – it is derived from the Malay language spoken along the Straits of Malacca and was originally the trade language of the region.

There is a growing interest in Indonesian NLP, with efforts in creating resources and tools such as morphological analysers (Uliniansyah *et al.* 2002; Asian *et al.* 2005; Pisceldo *et al.* 2008; Mistica *et al.* 2009; Larasati *et al.* 2011), developing corpora either as a balanced collection of texts representing different genres for linguistic investigation (Arka *et al.* 2009) or a large scale corpus for NLP tasks such as statistical machine translation (Riza *et al.* 2008).

In the late 2000s, there were few large-scale publicly available resources and tools for NLP in Indonesian for research purposes, with the aforementioned studies (Arka *et al.* 2009; Riza *et al.* 2008) on resource and corpus creation being at the initial planning phase at the time of publication. However, even at this time, it was difficult to define Indonesian as a 'low-density' language because by definition these languages have minimal web presence (Baldwin *et al.* 2006); Indonesian had and still has a large presence in the world wide web, with large news agencies such as Detik, Kompas, and even the BBC delivering language content online. It is however, under-resourced, but quantifying how under-resourced it actually is is difficult to determine.

It is beyond the scope of this study to provide an in depth survey of available resources for Indonesian NLP, but as a way to gauge how comparatively under-resourced Indonesian is with respect to other languages with the same worldwide speaker population, we gather data from two sources.[1] The first is Ethnologue[2] as a reliable source

[1]For some background on available Natural Language Processing resources for Indonesian see Section 6.1.1.

[2]`www.ethnologue.org`

for speaker population; the second source is the Linguistic Data Consortium (LDC)[3] with its catalogue of language resources to represent the bank of linguistic and NLP materials available for the language. We chose LDC because it represents a consortium of universities and institutions who provide data. Using the LDC is simply a way to approximate the kind of resources available for a given language, and we do not claim that the LDC is the definitive source for linguistic and NLP resources.

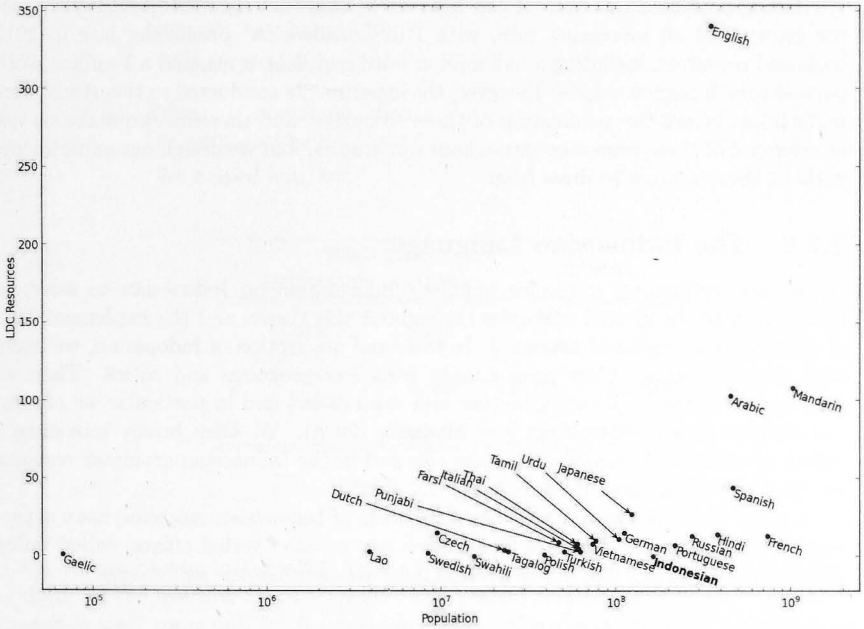


Figure 2.1: Worldwide speaker population plotted against number of resources found in LDC

We plot the data we had collected from LDC and Ethnologue simply as a way to gauge how languages compare. Although Indonesian is listed as one of the LDC languages, it unfortunately has no resources. The only other language shown in Figure 2.1 that also returned no resource was Swahili, which has a smaller worldwide

[3] `www.ldc.upenn.edu/`

speaker population than Indonesian has, with around 15.5 milllion speakers in comparison to the world wide speaker population of Indonesian, which today should be well in excess of 165 million based on data from the year 2000 (Gordon 2005).

Figure 2.1 shows that for the languages with a similar worldwide speaker population, Indonesian is at the bottom, with Japanese and German having a healthy number of resources. This figure also shows that the LDC specialises in the dissemination of English resources. Although this is not a complete picture of the resources for the languages, it serves to approximate how well resourced some languages are in comparison to others.

Having presented this data, it can however be said that Indonesian NLP resources are growing at an increasing rate, with Pan Localization[4] publishing late in 2012 tools and resources, including a half million word parallel corpus, and a 1 million word part-of-speech tagged corpus. However, the experiments conducted in this study were undertaken before the publication of these resources, and therefore we make no use or reference of these resources throughout our studies, and we design our experiments without these sources to draw from.

2.2.2 The Indonesian Language

In this section, we introduce linguistic information on Indonesian to assist in understanding the glossed examples throughout this thesis, and the implementation of the linguistic resources presented. In this brief description of Indonesian, we begin with a short outline of the pronominals, both free pronouns and clitics. Then we turn our attention to phrase structure and word order, and in particular we present the analysis presented by Arka and Manning (2008). We then briefly introduce a subset of affixes in Indonesian that are encoded in the Indonesian grammar resource we employ (see Chapter 4).

A more in depth examination of two aspects of Indonesian morphosyntax is presented in the next two sections, focusing on two kinds of verbal affixes, called **voice markers** in Section 2.2.3; in the Section 2.2.4 we examine the suffix *-kan*.

Pronominals

Indonesian pronouns are not marked for gender or case, however *ia* "he/she" is a special case, which Sneddon (1996) describes in terms of relative position in the clause, and which Musgrave (2001) defines functionally. It seems to mark agents, or at least the role that's highest on the thematic hierarchy (see Section 2.3.1 for a brief explanation of thematic hierarchy). However, there is no such restriction on *dia* "he/she", as can be seen in Example (2.1).

[4] `http://panl10n.net/english/OutputsIndonesia2.htm`

(2.1)

a. ***Dia*** *menolong kami.*
3SG AV-help 1PL.INCL
"(S)he helped us."

b. ***Ia*** *menolong kami.*
3SG AV-help 1PL.INCL
"(S)he helped us."

c. **Kami menolong* ***ia.***
1PL.INCL AV-help 3SG
"We helped him/her."

d. *Kami menolong* ***dia.***
1PL.INCL AV-help 3SG
"We helped him/her."

	SINGULAR		PLURAL
	INTIMATE	NEUTRAL	
1	*aku*	*saya*	*kita* (INCL); *kami* (EXCL)
2	*(eng)kau* *kamu*	*anda*	*kalian* (intimate)
3	–	*ia/dia*	*mereka*

Table 2.1: Pronouns

As can be seen in Table 2.1, for the first person plural pronouns, there is a different form for the inclusive (INCL – speaker and hearer) and exclusive (EXCL – speaker and non-hearer) pronoun. In the singular column, there is an INTIMATE and NEUTRAL distinction for singular pronouns for first and second person. With the exception of *kalian* (second person plural), all plural pronouns are neutral.[5]

In addition to the free pronouns, Indonesian also has pronominals clitics, which orthographically attach to the verb of which it is an argument. There are two kinds of

[5]The intimate and neutral distinction can be likened to the contrast with the Spanish *tú* and *ústed* which are the *familiar* and *non-familiar* second person pronouns, respectively. However, the usage of the Indonesian 'neutral' pronouns are more akin to the way in which the non-familar pronouns are used in South American Spanish, where they are in more common usage, than the Spanish spoken in Spain. Hearing *anda* "you" between adults who have known each other for a long time, and consider each other friends is not uncommon in Indonesian, especially if they are of the opposite sex. This is in slight contrast with the usage of *usted* "you" in Spanish.

pronominal clitics: prefixing and suffixing, which can roughly be described as agent and patient clitics, respectively. This is true with the exception of the third person clitics, as seen from Table 2.2.

	AGENT
1	*ku-*
2	*kau-*
3	*di-V-nya*

	PATIENT/POSS
1	*-ku*
2	*-mu*
3	*-nya*

Table 2.2: Agent and Patient/Possessive Clitics

(2.2) *Buku itu* ***ku****beli dan* ***ku****baca.*
book that 1sg=buy and 1sg=read
"I bought and read that book."

(2.3) a. *Saya* ***me****lihat****nya****.*
1sg AV-see=3sg
"I saw him/her/it."

b. *Buku itu* ***di****lihat****nya****.*
book this UV-see=3sg
"He/she saw the book."

Example (2.2) from Musgrave (2001) shows that these clitics are orthographically bound to their verbs like affixes. As for having their own syntactic position, we discuss this further in Section 4.2.2.

We also see an example of the third person pronominal clitic *-nya* in agent and patient roles.[6] However, there are two different verbal affixes, highlighted, which we have glossed as AV and UV here, called voice markers, which we discuss in Section 2.2.3.

For politeness, one avoids using pronouns in Indonesian when addressing people, and in such cases, pronoun substitutes can be used instead (Sneddon *et al.* 2010), as shown in Examples (2.4), and (2.5)

(2.4) *Ini untuk Dinah*
this to D
To Dinah: "This book is for you, Dinah."

[6]In addition, *-nya* is polysemously a marker of definiteness, as seen in the example:

Jam*nya* benar.
hour=DEF true
"That's the right time." Lit: "That hour is right."

(2.5) *Buku ini sudah Pak Arka baca.*
book this already Mr A read
To Pak Arka: "This book, you already read, Pak Arka."

Example (2.4) can be said in reference to Dinah to mean "This is for Dinah", however the construction shown in Example (2.5) is only used in direct address.

Phrase Structure and Word Order

In the same way that English determines the relationship of the arguments to their verb by their relative position in the phrase, so does Indonesian. This is in contrast with Tagalog, a language related to Indonesian where the order of the arguments in the clause does not alter the overall interpretation of the clause.

In this sense, Indonesian is a configurational language, meaning that positions in the phrase structure determines grammatical relations, as suggested by the phrase structure proposed by Arka and Manning (1998, 2008) in Figure 2.2.

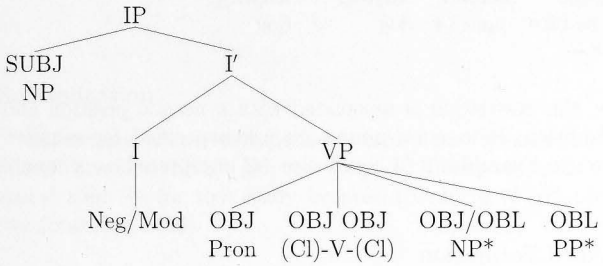


Figure 2.2: Phrase Structure per Arka and Manning (2008)

The phrase structure in Figure 2.2 suggests that neither enclitics nor proclitics have a phrase structure position, and that a separate syntactic node is required for preverbal pronouns. Conversely, Musgrave (2001:45) conservatively analyses both proclitics and enclitics as having syntactic positions.

In terms of simple word order, from this phrase structure tree, we can take the SUBJ(ect) as being the constituent to the left of the verb. The positions for the OBJ(ect) reside closest to the verb, and OBL(iques) appear to the rightmost of the clause. The Kleene star "*" represents zero or more of the constituents it appears after, while parentheses "( )" indicate optionality (i.e. one or zero occurences).

With respect to noun phrases, Porterfield and Srivastav (1988) observe that there are positional restrictions to definiteness for bare common nouns in Indonesian. Porterfield and Srivastav (1988), in Example (2.6) show that the bare noun phrase *pohon* "tree" cannot refer to a particular tree. However, with a determiner as shown in Example (2.7) below, this is not the case:

(2.6) *Saya melihat* ***pohon.***
1sg see tree
"I see a tree."/"*I see the tree."

(2.7) *Saya melihat* ***pohon itu.***
1sg see tree that
"I see the tree."/"*I see a tree."

On the other hand, Porterfield and Srivastav (1988) show, with the following active-passive example pairs, that the subject position cannot have a default indefinite interpretation for bare nouns.

(2.8) ***Seekor anjing*** */* ****anjing*** *menggigit kaki saya.*
one-CLF dog / dog bite leg 1sg
"A dog bit me."

(2.9) *Kaki saya digigit* ***seekor anjing*** */* ***anjing.***
leg 1sg PASS-bite one-CLF dog / dog
"A dog bit me."

More importantly, this restriction is associated with syntactic position and not thematic roles. In addition, Indonesian nouns are underspecified for number and therefore *pohon* "tree" in Example (2.6) could also be interpreted as a number of trees and not just one.

Morphology and Word Formation

Much of the description on Indonesian word formation is in terms of how to derive from one word class to another (Kridalaksana 1998; Sneddon 1996; Muhadjir 1981). This seems appropriate because, with the exception of reduplication, which we discuss in this section, it is claimed that Indonesian exhibits only derivational morphology (Musgrave 2001). Muhadjir (1981) studies how the order of affixation occurs when multiple affixes or morphological processes are combined in the variety of Indonesian spoken in Jakarta. In particular, he posits affixation order by the intermediate word forms that are allowable by the language.

The idea that the order in which morphological processes apply explains the difference between *memukul-mukul* "hitting" and *pukul-memukul* "hit each other". The stem on both verbs consists of *pukul* "hit" with the affix *meN-*, with the doubling of the stem. However the order in which they apply affects their surface form, as shown by Mistica *et al.* (2009).

Some common affixes are described in Table 2.3, and for the rest of this section we concentrate on reduplication, and discuss the affixes described as *voice* markers in Section 2.2.3, and the suffix *-kan* in Section 2.2.4.

AFFIX	GLOSS	DESCRIPTION
meN-	AV	These are known as **voice** markers
di-	PASS/UV	and they signal the alignment of arguments
∅-	AV./UV.	(see Section 2.2.3)
-kan		These have various functions, but are most
-i		commonly characterised as applicatives.
ber-	BER	Forms intransitive verbs.
ter-	TER	This prefix has been described as an accidental passive
pe...an	PE AN	
pe-	PE	These are types of nominalisers.
-an	AN	
se-	SE	has various functions e.g. *se* +noun = 'one noun'; *se* +adj = 'have the same adj'

Table 2.3: Common affixes

Reduplication

Indonesian has three types of reduplication: partial, imitative[7] and full reduplication (Sneddon 1996). We only consider full reduplication — or full repeat of the lexical stem — for this study because this form of reduplication is highly productive (Sneddon 1996).

(2.10) REDUPLICATION OF STEM

duduk-duduk	*sakit-sakit*
sit∼sit	sick∼sick
"sit around"	"be periodically sick"

(2.11) REDUPLICATION OF STEM WITH AFFIXES

memukul-mukul	*pukul-memukul*
meN-hit∼hit	*meN*-hit∼hit
"hitting"	"hit each other"

Reduplication seems to perform a number of different operations. There is the operation, which affects verbal aspect, affecting how the action is performed over time. These examples are seen in (2.10) *sakit-sakit* and (2.11) *memukul-mukul* are

[7]This is also known as rhyming jingles. See Pawley (2010) for the semantic variation of rhyming jingles.

comparable to the English progressive *-ing* in that the action is performed over a period of time.

2.2.3 Voice and Passive

This section describes Austronesian voice and briefly compares it with the passive. The similarities they have is that both contribute information that trigger a different association of thematic roles to argument functions.

The passive construction in English is a means of rearranging the grammatical status of the participants in an action, which also reduces the number of direct arguments required by the verb. For instance, Example (2.12a) shows a non-passive verb taking two arguments, with the agent in subject position. In Example (2.12b), we see the passive construction with only one obligatory argument – the patient in subject position and the agent being optional.

(2.12) a. The boy ate the cookie.
b. The cookie was eaten (by the boy).

Kibort considers the passive as a kind of subject-affecting operation, with the operation it undergoes preventing the most agentive participant from achieving "its normal surface realisation" (Kibort 2004:57) as the subject. The passive is a way to make the most prominent participant in the clause less prominent.

On the other hand **voice marking** which is realised as an affix on the verb, is described as signalling the thematic status of the subject (Musgrave 2008). The commonly cited example to showcase this phenomenon is from Tagalog, a language spoken on the island of Luzon in the Philippines.

(2.13) *b-**um**-ili* *ng* *isda* *sa* *tindahan* ***ang*** ***lalake***
VC-buy CORE fish OBL store man
"The man bought fish in the store."

(2.14) *bi-bilh-**in*** *ng* *lalake* *sa* *tindahan* ***ang*** ***isda***
IRR-buy-VC CORE man OBL store fish
"The man will buy the fish in the store."

(2.15) *bi-bilh-**an*** *ng* *lalake* *ng* *isda* ***ang*** ***tindahan***
IRR-buy-VC CORE man CORE fish store
"The man will buy fish in the store."

(2.16) ***ipam**-bi-bili* *ng* *lalake* *ng* *isda* ***ang*** ***salapi***
VC-IRR-buy CORE man CORE fish money
"The man will buy fish with the money."

(2.17) ***i**-bi-bili* *ng* *lalake* *ng* *isda* ***ang*** ***bata***
VC-IRR-buy CORE man CORE fish child
"The man will buy fish for the child."

The above examples from Foley (2008) show how the voice markers (VC) in each of the sentences (in bold) lead to a different participant being highlighted as the *ang*-marked phrase[8] (also in bold). Example (2.13) is the *actor voice* (AV) form of the verb, with the *um* affix, which dictates that the actor be the *ang*-marked constituent. The *undergoer voice* (UV), in Example (2.14) highlights the most effected participant, while Examples (2.15) – *locative voice*; (2.16) – *instrumental voice*; and (2.17) – *benefactive voice*, highlight the location, instrument, and beneficiary, respectively.

The term voice is applied to processes that realign thematic roles onto their grammatical functions (Jukes 2012). The effect the voice realignment has on the verb has been analysed in a number of ways in recent studies. They have been called **symmetrical voice** (Himmelmann 2005; Foley 2008; Arka *et al.* 2009), where the symmetrical view of voice does not take either one of the voice markers as being primary. The attachment of the voice marker gives instructions for linking to argument functions. Certain voice constructions have been treated as ergative constructions [9] (Manning 1996; Arka and Manning 2008) and system (Maclachlan 1996).

It is difficult to have a unified account of voice because the properties from language to language differ greatly, and do not show the complete voice paradigm presented above. Even within Tagalog, not all verbs fall neatly within this description as seen in (Ramos and Bautista 1986), where some undergoer voice marked verbs (objective voice in their nomenclature) exhibit the *an* affix (LV - locative voice marker) rather than the *in* affix (UV- undergoer voice), as shown in Foley's (2008) for the Tagalog for the verb *bili* "buy".

It can be seen with Puyuma, an Austronesian language native to Taiwan, that an analysis that is comparable with the Tagalog voice system can be made of its voice markers, as shown in Examples (2.18) to (2.21).

(2.18) *trakaw* *dra* *paisu* *i* *isaw*
<AV>steal ID.OBL money SG.NOM Isaw
"Isaw stole money."

[8]The *ang*-marked phrase has a special status in Tagalog. This is considered the pivot (grammatical subject) in LFG (Kroeger 1993).

[9]Terms such as *symmetrical*, *ergative*, *accusative* refer to alignment properties of the language, and rely on concepts defined by Dixon (1994) as S (sole argument of an intransitive verb), A (the most agentive argument of a transitive construction), O (the non-agent argument of a transitive construction). When S and A exhibit the same morpho-syntactic properties, this alignment is referred to as *nominative-accusative*, and when A and O display the same properties, this is called an *ergative-absolutive* alignment (See Dixon (1994) and Manning (1996)).

(2.19) *tu=trakaw-aw* *na* *paisu* *kan* *isaw*
3.GEN=steal-PV DF.NOM money SG.OBL Isaw
"Isaw stole the money."

(2.20) *tu=trakaw-ay=ku* *dra* *paisu* *kan* *isaw*
3.GEN=steal-LV=1SG.NOM ID.OBL money SG.OBL Isaw
"Isaw stole money from me."

(2.21) *tu=trakaw-anay* *i* *tinataw* *dra* *paisu*
3.GEN=steal-IV SG.NOM his.mother ID.OBL money
"He stole money for his mother."

These voice markers seemingly exhibit the same voice patterns as Tagalog, marking the thematic role of the subject, with Example (2.18) marking the *agent*, with the verb of the following example marking the *patient*, then the *locative* and then the *instrumental*. However, Teng (2008) analyses these verbal markers as markers of transitivity, rather than markers of prominence. Teng demonstrates that AV-marked verbs are intransitive, while the non-AV-marked verbs (of which there are three: patient, locative, and instrumental/benefactive) are transitive – to begin with, all AV-marked verbs lack the genitive object enclitic (*tu=*), suggesting their intransitive nature. In addition, non-subject arguments in these AV-marked constructions were shown to be obliques, from linguistic tests such as topicalisation and quantifier float (Teng 2008:150–155). Furthermore, she argues that if *em*[10] is analysed as a transitivity marker, or in this case a marker of intransitivity, rather than AV then this would provide a consistent account of this affix in Examples (2.18) to (2.21) from Teng (2008:161). This *em* affix, in Examples (2.22) and (2.23) is a marker that there is only one core argument, irrespective of its thematic role.

(2.22) ***m**-atel* *i* *drenan* *idri* ***na*** ***walak***
ITR-throw LOC mountain this.NOM DF.NOM child
"The child threw (something) away in the mountains."

(2.23) ***m**-atel* ***ku=paisu***
ITR-throw 1S.PSR=money
"My money was gone (disappeared)."

There seems to be no dedicated passive in Puyuma, but the prefix *ki* can emulate a passive-like construction, making the undergoer the subject, but there is no evidence that there is also a basic non-*ki*-marked transitive construction.

[10]What is seen as the suffix *m-* is shown to be an allomorph of *em* by Teng (2008:17)

(2.24) *ki-sulu-sulud=ku* *dra* *trau*
PASS-RED-push=1S.NOM ID.OBL person
"I got pushed by others."

The facts for Makassarese, a language spoken in South Sulawesi, Indonesia, give us a similar voice profile to Puyuma, except that there is a true passive in Makassarese, which is much like the Indonesian *di-* (Jukes 2012).

(2.25) *Nikokkoka'* *(ri* *meongku)*
nikokko'=a' (ri meong-ku)
PASSbite=1ABS (PREP cat-1.POSS)
"I was bitten (by my cat)"

(2.26) *Nakokkoka'* *meongku*
na=kokko'=a' meong-ku
3ERG=bite=1ABS cat-1.POSS
"My cat bit me."

Makassarese is not a symmetric language; it is an ergative language that has two kinds of actor voice markers. The first of these markers is **aC**,[11] which is attached to nouns and intransitive verb stems. The second of these markers is **aN**,[12] which can attach to transtive verbs. The patient in these constructions can be expressed or assumed. Like Puyuma, it is shown that these verbal prefixes are valence signalling rather than salience marking.

Indonesian differs again from Foley's (2008) symmetrical analysis of Tagalog, and the transitivity analysis of Makassarese and Puyuma. To begin with the AV marker does not exclusively mark transitive or intransitive verbs,[13] for example Examples (2.27) and (2.28) are both prefixed with the AV marker with the latter exhibiting one argument and the former two.

(2.27) [from Stevens and Schmidgall-Tellings (2004:653)]

Hasil panen menaik.
yield crop AV-climb
"The crop yield is increasing."

(2.28) *Ibu* *membeli* *daging* *bercenang.*
mother AV-buy meat chopped
"Mother bought chopped meat."

[11] **aC** = the open vowel 'a' followed by a 'C'onsonant.
[12] **aN** = the open vowel 'a' followed by a 'N'asal.
[13] In fact the *ber-* prefix is often described as intransitive verb affix.

Indonesian has an actor voice (AV), an undergoer (UV), as well as a passive (PASS). The passive is formed with the prefix *di-* as seen below in Example (2.30). The actor voice (AV) is signalled with a *meN-* prefix as shown in Example (2.29).

(2.29) *Amir* ***mem****baca* *buku* *itu*
Amir AV-read book this
"Amir read the book"

(2.30) *Buku* *itu* ***di****baca* *(oleh)* *Amir*
book that PASS-read by A
"That book was read by Amir."

Arka and Manning (2008) describe the undergoer voice construction that arises from the combination of the verbal prefix *di-* with the suffix *-nya*, where *-nya* is the 3rd person agent (shown as *di-V-nya* in Table 2.2).

(2.31) *Dirinya* *tidak* ***di****perhatikan****nya***
3REFL NEG UV-care=NYA
"(S)he didnt take care of himself/herself."

Arka and Manning (2008) show that Examples (2.31) and (2.30) are structurally quite different through evidence in reflexive binding.

(2.32) *?*Dirinya* *tidak* *diperhatikan* *Amir*
3REFL NEG di-care-KAN Amir
"Himself was not taken care of by Amir."

(2.33) *Amir*$_i$ *diperlihatkan* *Ayah*$_j$ *foto* *dirinya*$_{i/*j}$
Amir PASS-show-KAN father picture 3REFL
"Amir$_i$ was shown the picture of himself$_{i/*j}$ by father$_j$."

The binding evidence from the following examples shows that *-nya* from Example (2.31), is quite different from the agent Amir in Example (2.33).

Furthermore, Arka and Manning (2008) show that it cannot bind the reflexive subject, as seen in Example (2.32), nor can it bind the theme object, as seen in Example (2.33).

This demonstrates that the agent *=nya*, when the verb is suffixed with *di-*, should be analysed as having a different relationship to the verb than other agents

The passive is signalled with PASS, unless *-nya* is also encliticised with the 3rd person pronoun *=nya*, in which case we have the undergoer construction. Another characteristic of the passive is that the agent is not obligatory, and in Indonesian the preposition *oleh* "by" is also optional in an agent *by-phrase* if the agent is directly after the verb, as shown by the '()' in Example (2.30).

There is another undergoer construction that is signalled by word order, which has been labelled **Pro-V** by Musgrave (2001), and object preposing (or shifting) by Chung (1978), for example Example (2.34):

(2.34) [from (Chung 1978:335)]

Kejadian itu kita lihat kemarin
accident this we see yesterday
"The accident we saw yesterday."

This construction obligatorily has the agent directly to the left of the bare verb, hence the label Pro-V. Chung (1978:344) observes that this agent position is highly restricted:

> But in object preposing clauses, the underlying subject must be a pronoun or (less felicitously) proper noun; it is never a full NP

Characterising Indonesian voice markers as signalling valence rather than signalling whether the agent or undergoer should occupy the subject position is a little difficult. Through the topicalised construction, shown in Example (2.35), Arka and Manning (2008) demonstrate that for the AV-marked verb, the non-subject *-nya* is a term. In Indonesian, pronominal copy is only possible with core arguments according to Arka and Manning (2008:25), and in this example we see that this is possible:

(2.35) *Orang itu, saya yang menolongnya*
person that I REL AV-help=3sg
"As for the person, I helped him/her."

With respect to voice systems and Austronesian languages, Voskuil (2000:212) comments that:

> (o)ne would expect that within a language family, the superficial, outwardly observable properties remain relatively constant, but that the inner workings of these language can vastly differ.

Such is the case shown by Teng (2008) for Puyuma, but Indonesian seems to share more characteristics with Tagalog than Puyuma or Makasarese, due to how voice markers affect the alignment of arguments. However, Indonesian, like Makassarese and unlike Tagalog and Puyuma, does exhibit a passive.

Unlike the passive, which is a subject-affecting operation that prevents the highest ranked argument for achieving its normal surface realisation, Austronesian voice seems to do the opposite: the marking on the verb indicates which participant will achieve

subject status. This is the main reason for the use of the term 'symmetrical voice system' – there is no default alignment (i.e. neither an ergative nor an absolutive alignment is basic) but rather, the verb indicates this alignment of arguments and realisation.

In addition there is a split in the voice systems seen here, even with the four languages we have briefly looked at in this section, between languages where voice is a marker of transitivity (Puyuma and Makassarese) and those where it is not that clear (Indonesian and Tagalog). Although there are problems with the symmetrical view of voice, we adopt this concept for Indonesian in this thesis. For Indonesian, because there are only two voice types,[14] this indicates whether we have a nominative-accusative (AV) or absolutive-ergative (UV) alignment for a clause. One of the problems with this analysis is that bare verbs (the morphologically unmarked verbs) in both Tagalog and Indonesian usually feature in UV constructions, for instance object preposing or the Pro-V construction in Example (2.34). However, we do represent this symmetrical voice in the implementation in Chapter 4 by ensuring the obligatory VOICE feature for all clauses, with no default interpretation for alignment. Instead the surface realisation of the highest argument is always determined by the VOICE feature.

2.2.4 *-kan*: the profile of a deviant affix

In this section, we summarise the possible uses of *kan*, and then draw our attention to the two main constructions: the applicative benefactive, and causative. The behaviour *-kan* imposes has been attributed to the kind of stem involved in the *kan* construction in previous accounts, and these studies have been dedicated to characterising these stem classes (Dardjowidjojo 1971; Arka 1993; Vamarasi 1999). We end this section with an account of how these stems classes are determined. However, a problem with the stem classes account of the variation imposed by *kan*, as Kroeger (2007) points out, is that not only is there such a vast spectrum of possibilities, but also the same stem can result in different constructions when affixed with *-kan*. Kroeger (2007) accounts for the varying outcomes of *-kan* as homophony. That is, there is more than one suffix *-kan* that accounts for the linguistic facts. However, despite giving the appearance of being a number of homophonous morphemes, Son and Cole (2008) argue against this, claiming that although the suffix has more than one independent function, they are all related; they all involve causative semantics, and "the aspectual meaning of which involves a causing event and a caused eventuality (or a result state)" (Son and Cole 2008:121). One thing that is certain, however, is that the distribution of *-kan* is not straightforward or clear cut.

[14]This excludes the passive.

A Spectrum of Variation

Kroeger (2007) and Son and Cole (2008) describe a spectrum of morphosyntactic and semantic variation that *-kan* imposes; it has been described as a **transitiviser** – these valence increasing processes are most commonly characterised as **applicativisation**, allowing a benefactive to act as a core argument as shown in Example (2.36a) (with the non-*kan*-affixed verb shown in Example (2.37)) or **causativising**, seen in Example (2.36b).

(2.36) [from (Kroeger 2007)]

a. *Ibu* *menjahitkan* *saya* *baju.*
mother AV-sew-KAN 1sgshirt
"Mother sewed me a shirt."

b. *Saya* *menjahitkan* *baju* *ke* *tailor.*
1sgAV-sew-KAN shirt to tailor
"I had my shirt sewn by a tailor."

(2.37) [from Wikipedia page *April 2007* under *Rabu, 11 April 2007*]

dan *dokter* *menjahit* *luka-lukanya*
and doctor AV-sew wound∼wound=3sg
"and the doctors sewed his/her wound"

Kroeger (2007) also points out that in some contexts it does not increase valency at all, but instead **changes the semantic role** of the direct object, as shown in Example (2.38).

(2.38) [from (Kroeger 2007)]

a. *Perawat* *membalut* *lukanya* *dengan* *kain.*
nurse AV-wrap wound=3sg with cloth
"The nurse wrapped his wound with a bandage."

b. *Perawat* *membalutkan* *kain* *ke* *lukanya.*
nurse AV-wrap-KAN cloth to wound-3sg
"The nurse wrapped a bandage around his wound."

This example is different to Son and Cole's (2008) observation, who note that for the GOAL-PP *kan* constructions, as seen in Example (2.39), the suffix *-kan* is **optional**.

(2.39) [from (Son and Cole 2008)]

a. *Dia mengikat tali itu.*
3sg AV-tie rope that
"S/he tied the rope."

b. *Dia mengikat(-kan) tali itu ke anjing*
3sg AV-tie-KAN rope that to dog
"S/he ties the rope to the dog."

Example (2.38) shows that the instrument occupying the object position is directly after the verb, while in Example (2.39b) the predicate argument structure of *mengikat* "tie" does not change with the affixing of *-kan*, rendering it optional. Son and Cole (2008:132) note that in the previous literature[15] these resulting changes to the argument structure have been labelled the **instrumental** *-kan* construction. Another example from Kaswanti (1997) shows that this instrumental usage of *-kan* can also alter the predicate argument structure of the resulting verb, as we see in Example (2.40). This example seems to parallel Kroeger's (2007), shown in Example (2.38).

(2.40) [from (Kaswanti 1997)]

a. *John menikam perut harimau dengan belati.*
J AV-stab belly tiger with dagger
"John stabbed the tiger's belly with a dagger."

b. *John menikamkan belati ke perut harimau*
J AV-stab-KAN dagger to belly tiger
"John stuck the dagger into the tiger's belly."

Another usage of this suffix has been labelled the **locative alternation**, because it alternates with the locative suffix. Verbs bearing the *-i* suffix indicate that the direct object is a location, while these same verbs with the *-kan* suffix have as their direct object a displaced theme (Kroeger 2007; Arka *et al.* 2009). An example of this is shown in Example (2.41). An example of the non-suffixed verb in Example (2.42) for comparison.

(2.41) [from (Kroeger 2007) and (Arka *et al.* 2009)]

a. *Buruh itu memuat**kan** beras ke kapal.*
worker that AV-hold-KAN rice to ship
"Workers loaded rice onto the ship."

[15]For example Son and Cole (2008) refers to Arka (1993); Sneddon (1996); and Postman (2002)

b. *Buruh itu memuati kapal dengan beras.*
worker that AV-hold-I ship with rice
"Workers loaded the ship with rice."

(2.42) [from Wikipedia article *Kamus Besar Bahasa Indonesia Pusat Bahasa*]

Kamus edisi ketiga ini memuat sekitar 78.000 lema.
dictionary edition third this AV-hold around 78,000 lemma
"The third edition holds around 78,000 lemmas."

In Example (2.41a), the direct object of the *kan*-suffixed verb is the 'theme' ("rice", which is being loaded onto the ship), whereas Example (2.41b) with the *i*-suffixed verb, the direct object is the location ("ship", where the rice is being loaded).

Kroeger (2007) observes that if we compare examples such as Examples (2.39) and (2.40) with the causative in Example (2.36b), they appear to be a variation of the causative use of *-kan* rather than a dedicated instrumental construction. Likewise, Kroeger (2007) shows that the locative alternation can indeed be semantically decomposed as another variant of the causative.

It has also been observed that the suffix can seemingly reduce the valency of a verb. For example, the pattern seen in Example (2.43) seems to reduce the valency of the verb *beri* "give" when affixed with *kan*. In Example (2.43a), there are two direct arguments, but in Example (2.43b) the form of the complements are NP + PP.

(2.43) [from (Kroeger 2007)]

a. *John memberi Mary buku itu.*
J AV-give M book that
"John gave Mary the book."

b. *John memberikan buku itu kepada Mary.*
J AV-give-KAN book that to M
"John gave the book to Mary."

(2.44) [from (Son and Cole 2008)]

John memberi(**kan**) surat itu kepada Peter.*
J AV-give-KAN letter that to P
"John gave the letter to Peter."

Son and Cole (2008) note that the pattern exhibited in Example (2.43) is due to the stem being inherently ditransitive, and in fact the suffix *kan* is obligatory when the complements of the verb are of the form NP + PP, as shown in Example (2.44). Given these kinds of examples, Son and Cole (2008) argue against the analysis of *-kan* being syntactically a transitiviser.

In addition to these, Arka *et al.* (2009) attribute the **comitative** reading of Example (2.45b) to the affixing of *-kan* to the verb *datang* "come". This example shows two different readings for the *kan*-affixed verb. Arka *et al.* (2009) call reading (1) from Example (2.45b) the comitative-applicative *-kan*, and (2) the causative *-kan*.

(2.45) [from (Arka *et al.* 2009)]

a. *Polisi datang.*
police come/arrive
"The police arrived/came."

b. *Mereka mendatangkan polisi.*
3pl AV-come-KAN police
"They arrived with the police" (1)
"The called for/made the police come" (2)

The suffix *-kan* exhibits a large variety of uses; it has shown to introduce an applicative benefactive object, form causative constructions, express a displaced theme or instrument as the direct object as a variation of the causative, enable a comitative-applicative reading, seemingly reduce valency, shift the relative prominence of the complements, as well as exhibit no discernable change on the verb, rendering it optional on certain stems. Before presenting some ways linguists have characterised and explained these variations attributed to *kan*, we delve into the some of the features of *-kan* in the benefactive applicative construction, and then causative uses outlined by Arka (1993).

Applicative

An applicative construction is a means by which a thematically peripheral argument or adjunct can be encoded as a core object argument (Peterson 2007). According to Peterson (2007), the two ways that we can describe applicativising languages are: symmetrical and non-symmetrical. A symmetrical language treats both objects, the original object and the introduced applicative, in the same way, allowing syntactic operations, such as relativisation and passivisation to be performed on either. Most languages sit in between the extreme cases of the symmetrical/non-symmetrical classification.

Chung (1976:42) describes *-kan* having two basic effects:

(i) it is an *object-creating rule*, in the sense that it turns the indirect object/benefactive into a direct object, and (ii) it displaces the underlying direct object so that it is inaccessible to later syntactic rules.

Also, Vamarasi's (1999) example shows that the benefactive NP in a *kan* construction is the only element that can be passivised, as shown in Example (2.46).

(2.46) [from Vamarasi (1999:74)]

a. *Teman saya dimasakkan nasi gorengnya.*
friend 1sg PASS-cook-KAN rice fried=DEF
"My friend was cooked the fried rice."

b. **Nasi gorengnya dimasakkan saya oleh Ibu.*
rice fried=DEF PASS-cook-KAN 1sgby mother
"The fried rice was cooked (for) me by Mother."

From this description, we would characterise Indonesian as falling into the non-symmetrical category of applicativising languages.

However, surprisingly, we find many examples in Wikipedia, where either the theme or benefactive object, in this kind of double object construction could be relativised. For example, either of the objects in the *-kan* double object construction for the verb *beli* "buy" can be relativised, as we see in Examples (2.47) and (2.48).

(2.47) [from Wikipedia page *George Gershwin*]

Kakak George yang bernama Ira dibelikan piano
brother G REL BER-name I PASS-buy-KAN piano
"George's brother whose name is Ira was bought a piano"

(2.48) [from Wikipedia page *Vanessa-Mae*]

Biola Guadagnini dibelikan oleh orang tuanya
Violin G PASS-buy-KAN by CLF parent+3sg
"The Guadagnini violin was bought by her parent."

We also found verbs like *beri* "give", which Son and Cole (2008) claim to be inherently ditransitive, and allows passivisation on either object as shown in Examples (2.49) and (2.50).

(2.49) [from Wikipedia page *Malaysia*]

Sebagian besar orang Malaysia diberikan kewarganegaraan oleh
most large CLF Malaysia PASS-give-KAN citizenship by
lex soli.
lex soli
"Most Malaysians are given/granted citizenship by lex soli."

(2.50) [from Wikipedia page *Selandia Baru*]

Nama ini diberikan oleh Abel Tasman seorang
name this PASS-give-KAN by A T one-CLFexplorer
penjelajah dari Belanda.
from Netherlands
"This name was given by Abel Tasman an explorer from The Netherlands."

There were also examples where either object can be clefted in a *yang* construction. An example of each variation is shown in Examples (2.51) and (2.52).

(2.51) [from Wikipedia page *MS-DOS*]

Hanya IBM yang diberikan keleluasaan untuk terus
only IBM REL PASS-give-KAN opportunity to continue
menggunakan nama IBM PC-DOS, bukannya MS-DOS.
AV-use-KAN name IBM PC-DOS, not.actually MS-DOS
"Only IBM was given permission to continue using the name IBM PC-DOS, not MS-DOS."

(2.52) [from Wikipedia page *Suzi Quatro*]

Gitar bass pertamanya bermerek Fender Precision 1957 yang
guitar bass first=3sg BER-brand.name F P 1957 REL
dibelikan oleh sang ayah.
PASS-buy-KAN by honorific father
"His/her first Fender Precision bass guitar was bought by his/her father."

Although either object can appear as subject in a passivised construction, Example (2.47) allows the thematic object to remain adjacent to the verb *dibelikan* "was bought for", but the benefactive object in Example (2.46b) cannot occupy the same position, when the thematic object undergoes passivisation. This may indeed be the

reason why Vamarasi's (1999) Example (2.46b) is ungrammatical – the benefactive cannot appear directly after the passive verb.

Even though we find many non-elicited examples of the symmetrical nature of applicatives in Indonesian, there are many examples in the literature that support the claim that a thematic object occupying the subject position in a passive *kan* construction is ungrammatical. For example Kaswanti (1997) shows that passivising *buku* "book" in Example (2.53) is unacceptable.

(2.53) [from Kaswanti (1997:241)]

a. *John membelikan Mary buku itu.*
J AV-buy-KAN M book that
"John bought Mary the book."

b. *Mary dibelikan buku itu oleh John.*
M PASS-buy-KAN book that by J
"Mary was bought that book by John."

c. **Buku itu dibelikan Mary oleh John.*
book that PASS-buy-KAN M by J
"The book was bought for Mary by John."

Also, Kaswanti (1997) shows that there is some speaker variation in the usage of *-kan*, which accounts for the grammatical acceptance of Example (2.54); ordinarily *memberi* "give" would be affixed with *-kan* in such an example. Kaswanti (1997:235) calls this omission of the suffix *kan* an example of 'deviant' usage, through its non-usage.

(2.54) [*Deviant* example from Kaswanti (1997)]

John memberi buku itu kepada Mary
J AV-give-KAN book that to M
"John gave the book to Mary."

Although the passivisation of the thematic object has been shown in the literature to be ungrammatical, much to our surprise we found the passive construction with the theme-as-subject being used freely in Wikipedia.

There are a number of reason why we have found this variation of usages for the passive. It could be possible that the contributors of these pages are Indonesian speakers whose first language is a local language that allows symmetrical passivisation of applicativised double object constructions. Indonesian is the national language but there are many local languages that are spoken in the regions of Indonesian, with Indonesian being taught to children once they begin their schooling. One such Indonesian language where either the benefactive or the theme in an applicativised passivised construction is possible is Balinese (Arka 2008).

(2.55) a. *Sabilang anak cenik$_i$ beli-ang-a teken bapan-ne$_i$ tas*
every person child buy-APPL-PASS by father-3.POSSbag
"A bag was bought for every child$_i$ by his$_i$ father."

b. *Sabilang anak alit$_i$ ka-ambil-ang ajengan antuk*
every person small PASS-take-APPLfood by parents
reraman ipun$_i$-e
3.POSS-DEF
"Food was taken for every child$_i$ by his$_i$ parents."

Arka (2008) shows that there are two passives in Balinese. The difference between the *-a* passive in Example (2.55a) and the *ka-* passive in Example (2.55b) is the *a* passive is constrained to have a third-person actor (Arka and Manning 2008:80)[16]. One possibility for this widespread usage is that features of the local languages are seeping in Indonesian, allowing a more symmetrical usage, as we see with the examples we had found with respect to relativisation, the *yang* focus construction, and passivisation. As pointed out by Keller (2001), these examples also show the difficulty in relying solely on elicited binary grammaticality judgements.

Causative

Arka (1993) describes the causative uses of *-kan*, showing that the suffix can be concatenated with almost all grammatical categories, both open and closed. We see examples of meaning changes imposed by *-kan* in Table 2.4. This table shows a large range of semantic differences with the usage of the causative affix *-kan* between groups of bases with different parts of speech. It also shows that there is variation amongst word classes, with a defining separation within the verb class depending on its transitivity status. There is a much larger variation of causative meaning within the noun class from their glosses, and given these meaning changes for all stems/bases we see, we would assume a variety of syntactic expressions, just with the use of the causative *-kan*.

Arka (1993) also compares this morphological causative construction with a periphrastic causative, formed with the verb *buat* "make", seen in Example (2.56).

(2.56) [from (Arka 1993)]

a. *Ia menjatuhkan orang itu ke sumur.*
3sgAV-fall-KAN person that to well
"He dropped/threw the man into the well."
"He made the man fall down the well."

[16]Arka (2008:81–82) demonstrates the oblique status of the agent phrase, and therefore they are not core direct arguments.

(1)	Base=V			
	(i)	Intransitive verbs		
		terbang "fly"	*terbang-kan*	"cause to fly"
		masuk "enter"	*masuk-kan*	"cause to enter"
	(ii)	Transitive verbs		
		jahit "sew"	*jahit-kan*	"have something sewn by X"
		ketik "type"	*ketik-kan*	"have something typed by X"
(2)	Base=N			
	(i)	*arang* "charcoal"	*arang-kan*	"make X become charcoal"
		budak "slave"	*budak-kan*	"make X become a slave"
	(ii)	*budak* "slave"	*budak-kan*	"treat X as a slave"
		raja "king"	*raja-kan*	"treat X as a king"
	(iii)	*darat* "land"	*darat-kan*	"cause X to (move to) land"
		udara "air"	*udara-kan*	"cause X to (go to) land"
	(iv)	*penjara* "jail"	*penjara-kan*	"put X in jail"
		botol "bottle"	*botol-kan*	"put into a bottle"
	(v)	*obat* "medicine"	*obat-kan*	"have X treated medically by Y"
(3)	Base=A			
		besar "big"	*besar-kan*	"make X big"
		patah "broken"	*patah-kan*	"make X broken" / "break X"
(4)	Base=P			
		ke atas "up"	*keatas-kan*	"lift up"
		ke belakang "back"	*kebelakang-kan*	"move X to the back"
(5)	Base=NUM			
		satu "one"	*satu-kan*	"cause to become one" / "unite"
		dua "two"	*dua-kan*	"treat X as two"
		three "three"	**tiga-kan*	
(6)	Base=ADV			
		sangat "very"	*sangat-kan*	"make much more..."

Table 2.4: Examples of the application of *-kan* from Arka (1993:90)

b. *Ia* *membuat* *orang* *itu* *jatuh* *ke* *sumur.*
3sgAV-make person that fall to well
"He made the man fall down the well."

Arka (1993) shows that although the morphological causative can be equivalent in meaning with the periphrastic causative, there are subtle differences. While the periphrastic causative can be ambiguously interpreted as two causally related events that can be separated out in time, and therefore individually modified, the morphological causative encodes an event where the "causing and caused event cannot be singled out" (Arka 1993:96).

We have discussed both the applicative and causative nature of *-kan*. Arka (1993) also shows the applicativising nature of *kan*, assuming that applicativisation introduces a new argument, as per Bresnan and Zaenen (1990). Arka (1993), like Kroeger (2007), hypothesises two homophonous *-kan* affixes to account for all of the variations seen with *kan*-affixed verbs. Kroeger (2007) explicitly labels these homophonous affixes as KAN_1 and KAN_2. While Arka (1993) explains the semantics and the corresponding change in syntactic encoding of each, Kroeger (2007) views KAN_1 as modifying the semantic structure of the verb, with KAN_2 involving a change in the syntactic expression of the arguments, allowing the peripheral benefactive argument to be expressed as a direct object. The affix KAN_1 is underspecified in terms of the changes in the syntactic encoding of arguments, unlike KAN_2, but imposes a semantic change of a causative nature, which involves the Jackendoff-style lexical conceptual semantic expression CAUSE-BECOME-AT in the resulting verb (Jackendoff 1972; Jackendoff 2010). In this model all benefactive readings of *-kan* are attributed to KAN_1 and all non-benefactive readings of *-kan* involve the usage of KAN_2.

Hypothesising -kan

Kroeger's (2007) hypothesis captures many of the variations *-kan* exhibits, for example, it explains that the instrumental and so-called locative alternation uses are an application of KAN_2 (the syntactically underspecified semantic causative), and it also accounts for the optionality of *-kan* for some verbs, which is explained by their lexical heads already encoding a causative meaning, such as 'throw', 'send', 'pour'. However, Son and Cole (2008) claims that the spectrum of variation with the usage of *-kan* need not be accounted for by positing multiple homophonous affixes. They claim that the optional usage of the suffix as well as its causative and benefactive uses are not examples of accidental homophony; their thesis is that the suffix *-kan* is a morphological reflex of the RESULT head. This is a node that is projected as a result of the suffix *-kan* that introduces an RP (Result Phrase) which encodes a result state, which all *-kan* constructions share.

While Arka (1993); Kroeger (2007); and Son and Cole (2008) investigate the kinds of stems that *-kan* affixes in order to capture a better characterisation of the suffix, Dardjowidjojo (1971); Chung (1976); and Vamarasi (1999) aim to classify the kinds of stems that can host *-kan*. Dardjowidjojo (1971) takes morphological pattern based approach in his categorisation, investigating stems according to combinatorial possibilities; Vamarasi (1999) takes a lexical semantic approach, while Chung (1976) finds morphosyntactic classes that align with the variations attributed to *-kan*.

Dardjowidjojo (1971) identifies 7 subsets of verbs according to their morphological affixes, or the morphological patterns that they exhibit, focusing only on the prefix *meN*, and the two suffixes *-i* and *-kan* and their combinations. The 7 subsets he identifies are presented in Figure 2.3.

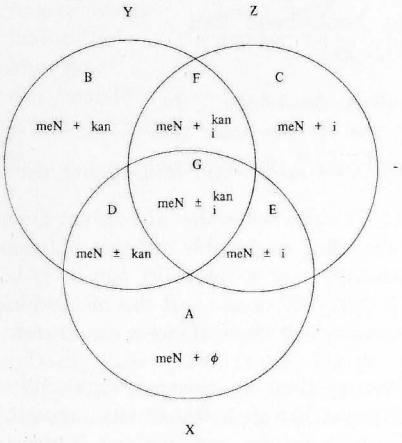


Figure 2.3: Dardjowidjojo's (1971) 7 stem types according to allowable affix combinations

Stems are categorised as belonging to these subsets according to their combinatorial possibilities. He then describes the semantic and syntactic variation he observes within these 7 groups. That is, the stems are manually classified as belonging to one of these subsets solely based on possible affixation. Therefore, stems can and do belong to the same subset even if their syntactic behaviour is not identical. For example, the stem *jauh* "far" and the verb *jual* "sell" are both described as belonging to subset G, even though *menjauh* "stay away" takes one direct argument, while *menjual* "sell" takes two, as seen in Examples (2.57) and (2.58).[17]

[17]The examples shown in Dardjowidjojo (1971) use the old orthographic conventions, which we change in these examples to reflect the modern Indonesian spelling after the spelling reforms in 1972.

(2.57) [from (Dardjowidjojo 1971:79)]

a. *Saya harus menjauh.*
1sgmust AV-far
"I must stay away."

b. *Saya harus menjauhkan dia.*
1sgmust AV-far him/her
"I must keep him/her away."

(2.58)

a. *Saya menjual buku-buku itu.*
1sg AV-sell book∼book that
"I sold those books."

b. *Saya menjualkan buku-buku itu.*
1sg AV-sell-KAN book∼book that
"He sold those books." (for someone)

Dardjowidjojo (1971) also notes that the subset D can be divided again into at least 5 subgroups according to possible different outcomes in transitivity when the stem (that can optionally host a the suffix *kan* or *i*) is affixed or not. From this study Dardjowidjojo (1971:83) deems that this method has inconclusive results in his endeavour to determine which kinds of stems can or must take which affixes:

> While to a certain extent we can see a regularity of the interrelation of the verbs with respect to their transitivity properties, it is found that no useful generalization can be made without having to add an open list of exceptions. This is apparently due to the fact that the co-occurrence with the base and the affixes is morphemically conditioned. Therefore, given a base, there is no way of telling what particular affix or set of affixes this base can or must take, and in many cases we do not know what kind of transitivity the resultant verb will acquire.

This study by Dardjowidjojo (1971) on changes to predicate-argument structure with the affixing of *-kan* (as well as *-i*), which was published almost two decades before Levin's (1989) study on the correlation of lexical semantics and syntactic alternation, showed that simply looking at a handful of examples, without the aid of corpus linguistic methods, and grouping them on their possible combinatorics did not find coherent groups of stems.

Sleeping Class	Working Class
tidur "sleep"	*kerja* "work"
masuk "enter"	*gurau* "joke"
timbul "arise"	*dusta* "lie"
mati "die"	*pikir* "think"
sampai "arrive"	*doa* "pray"
pulang "go home"	*nyanyi* "sing"
tenggelam "sink"	*bohong* "lie"
runtuh "collapse"	*batuk* "cough"
hilang "gone"	*main* "play"
jadi "make"	*cerai* "separate"
jalan "go"	*bicara* "speak"
renang "swim"	*tanya* "ask"
baring "lie"	*gambar* "draw"
alir "flow"	
diri "stand"	
seberang "cross"	

Table 2.5: Stems and their translations taken from Vamarasi's (1999) intranstive dichotomy

Vamarasi (1999) on the other hand takes a syntactico-semantic approach, rather than the kind of approach Dardjowidjojo (1971) takes, who groups verbs based only on morphological patterns. She claims that there are two kinds of intransitive stems that host the *-kan* suffix, which she calls the *Working Class* and *Sleeping Class* because *kerja* "work" and *tidur* "sleep" are representative verbs for each group, respectively.

Vamarasi (1999) shows that the morphosyntactic behaviour of the items in each class in Table 2.5 is the same. For example stems from the *Working Class* are never unaffixed when used intransitively; they are either prefixed with *ber-* or *meN-*. Stems from the *Sleeping Class* can remain unaffixed. More importantly these verbs when affixed with *-kan* gain the meaning of "X makes/causes/lets Y to Verb" according to Vamarasi (1999:28), while *Working Class* verbs do not. Not only do these *Sleeping Class* stems form causatives when affixed with *-kan*, but when they are not affixed with *kan*, Vamarasi (1999:29) claims that these are "unaccusative" verbs, while the *Working Class* verbs that are not affixed with *-kan* are "unergatives".

Chung (1976) also categorises bases according to their morphosyntactic behaviour, but does not make any generalisations about the semantic relatedness of the members of the class. There are three classes in which she categorises stems according to how they behave before and after the Dative rule has applied.The Dative, Chung (1976)

describes as either obligatorily affixing *-kan* or *-i* or having no affix at all. Chung (1976:56) describes three classes that alternate in the same way before and after the Dative is applied, as shown in Figure 2.4.

	Class I	Class II	Class III
Before Dative	-∅	-kan/-∅	-kan/-∅
After Dative	-kan	-i	-∅

Figure 2.4: Chung's (1976) stem classes according to allowable affixes in the Dative Alternation

Dardjowidjojo (1971), Vamarasi (1999) and Chung (1976) explain the behaviour of these valence changing rules, through the classification of *kan*-affixed stems. Kroeger (2007) points out that one problem with such a means of explaining the behaviour of *-kan* in this way is that many MEN+stem+KAN verbs are ambiguous, and particularly for Vamarasi's (1999) model, could result in multiple membership of her classes. For example Vamarasi (1999) claims that unaccusative intransitives in her model produces causatives when affixed with *-kan*, however the same *jahit* "sew", as we saw earlier in Example (2.36), can produce a causative and a benefactive meaning.

For decades, linguists have tried to describe and explain the idiosyncracies of the morphosyntactic product of *-kan*, however its exact nature is still unclear. In the early years of the investigation into *kan*, Dardjowidjojo (1971) and Vamarasi (1999) did this in terms of grouping like stems, with the former looking more at the surface word, and the latter the semantics of the stems. Kroeger (2007) has shown the problem with attributing the resulting behaviour of *kan* by attributing them to lexical categories alone, and therefore has hypothesised that this range of behaviour is due to there being two homophonous affixes KAN_1 and KAN_2. One applies morphosemantic changes pertaining to the causative, while the other is imposes morphosyntactic changes used in applicative constructions. Son and Cole (2008:121) also note with its range of different construction, that it gives this appearance, but argue against accidental homophony, and attributes the range of constructions with imposed by *-kan* as being related through verbal aspect.

2.3 Linguistic Theory

This section serves to introduce the basic linguistic assumptions made, and the theoretical underpinnings that inform the implementation in the development of the language resources.

Level of structure	**Type of linguistic information**	**Form of representation**
constituent structure	surface syntactic representation	tree diagram
functional structure	abstract GFs and features	attribute value matrix AVM
argument structure	valency	ordered list

Table 2.6: The parallel levels of representation, adapted from Mycock (2006)

2.3.1 Grammar Formalism – LFG

Lexical Functional Grammar (LFG) is a constraint-based formalism and lexically driven theory of syntax that has multiple distinct, but parallel, levels of representation, which enables the capturing of cross-linguistic variation, as well as similarities. Word order and constituency are described via rewrite rules, and grammatical information is encoded in the form of attributes and values.

The framework minimally consists of two levels or representation: *constituent-structure* and *functional-structure* (Kaplan and Bresnan 1982). There have been additional levels of representations proposed, such as *information structure*, *phonological structure*, *morphological structure*, and *semantic structure*. However, we only describe briefly the levels of representations summarised in Table 2.6 in the following paragraphs.

LFG is a formalism that employs unification operations. The constraints are spelled out in the functional structure (described below) that indicates the grammatical features. Unification is the process of combining this grammatical information, which is often derived from different parts of an expression.

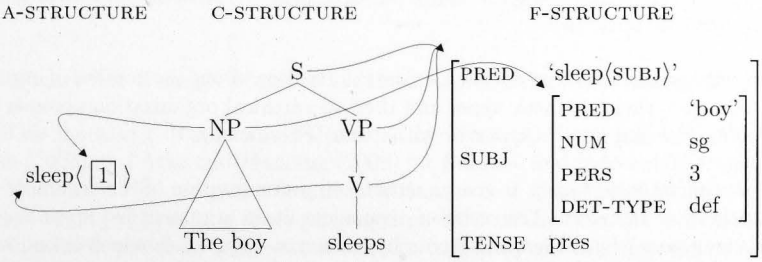


Figure 2.5: English sentence *"The boy sleeps."*

F(unctional)-structure represents linguistic information as attribute value matrices, as can be seen in Figure 2.5. One type of information expressed in these structures is dependencies in the form of grammatical relations (functions) listed such as *subject* (SUBJ), *object* (OBJ), and *oblique* (OBL), as well as adjuncts (ADJ), head-modifier relationships (MOD), and long distance dependencies (represented as coindexation).

This level of representation also exhibits grammatical features of the language (e.g. NUM, PERS, GENDER, and CASE). These features along with the grammatical functions are the constraints in the unification process in establishing verified analyses.

C(onstituent)-structure encodes syntactic information pertaining primarily to constituency, and the information regarding linear ordering of linguistic elements and their hierarchical organisation (i.e. phrasal structure) is represented as a tree, as seen in Figure 2.5. We point the reader to Dalrymple's (2001) chapter on *Constituent Structure*, which justifies the need for this level of representation. Within LFG, it is assumed that there is no information regarding the functions of arguments, such as SUBJ or OBJ, within the c-structure. It is this feature of LFG that provides flexibility to account for typologically diverse languages, which lends itself to such studies (Mycock 2006).

At this level of representation, we use symbols that represent the language's inventory of lexical categories, such as N(oun), P(reposition), V(erb), A(djective), ADV(erb), which can form the heads of phrases, as well as other categories that do not head phrases such as the particle *up* in *look an address* ***up*** in English (Dalrymple 2001).

X-bar theory is employed in LFG as a means of reflecting structural relationships between the internal nodes of the c-structure, and bounding the immediate sharing of grammatical features and their propagation. However it is not strictly or obligatorily used in implementation for grammar engineering purposes (see Dalrymple (2001) and Falk (2001) for an in depth discussion on how X-bar is used within LFG proper).

A(rgument)-structure like the f-structure and c-structure, is a parallel, independent level of representation. It can be thought of as the resolving of lexical semantics into syntactic structure.

> Argument structure encodes lexical information about the number of arguments, their syntactic type, and their hierarchical organization necessary for the mapping to syntactic structure. (Bresnan 1995)

But more importantly, it gives a structured representation of the argument-taking properties of the verb. The order of arguments in an argument list specified by the verb is imposed by a thematic hierarchy, shown in Figure 2.6, which is one way the prominence relations between the arguments are encoded.

agent > *beneficiary* > *goal/experiencer* > *instrument* > *patient/theme* > *location*

Figure 2.6: Thematic Hierarchy

The order of these arguments, according to Bresnan and Kanerva (1989), indicates their status with respect to the predicate and the order in which these arguments are composed with the predicate; roles that are lower down in the hierarchy are considered 'inner' arguments and in a sense more intrinsic to the semantics of the predicate based on the observation that these lower roles tend to be lexicalised rather than higher ranked roles.

The thematic hierarchy shown in Figure 2.6 designates the order in which the arguments appear (Bresnan 2001). If a predicate requires two arguments with the roles *agent* and *patient*, then the agent will precede the patient in the a-structure, as shown for the verb *hit* in Figure 2.7. The variation of argument structure we ascribe to is outlined by Manning (1996), where its realisation is syntactic rather semantic. In addition valence changing operations are performed at this syntactic level.

hit⟨ [1] , [2] ⟩ [1] Sarah hit [2] John.

Figure 2.7: Argument structure for *hit*

Furthermore, the resulting argument structures from these valence changing operations are nested argument structures (Manning 1996:43), as seen in the *causative* example in Figure 2.8. This a-structure represents the sentence *I melted the chocolate.*

CAUSE⟨ –, –, melt⟨ – ⟩ ⟩

Figure 2.8: Example of nested argument structure as per Manning (1996)

This modelling of a-structure described by Manning (1996) was also adopted in Arka (2003) and Arka and Manning (2008) for Balinese and Indonesian, respectively.

Manning (1996) represents argument structure as a hierarchical list where higher ranked arguments are in a more privileged position relative to other arguments that are in a lower position. This asymmetrical relationship governs certain grammatical constructions such as reflexive binding. Therefore in Indonesian, reflexive binding is not subject to the syntactic position or prominence of grammatical relations or syntactic position, but rules governing argument-structure.

This is also true of other Philippine-type languages, namely Tagalog and Cebuano, that allow a constituent that is syntactically less prominent to license a reflexive pronoun (Andrews 1985; Kroeger 1993). These binding facts have been explained through restrictions imposed by the thematic hierarchy (Jackendoff 1972; Kroeger 1993), and Manning's theory of argument structure formalises this observation as part of the theoretical framework. The flexibility and explanatory power of the syntacticised argument structure is due to the overloaded term *subject* having their overloading functions identified and distilled into separate components, as shown in Figure 2.9.

A-SUBJECT	the highest ranked argument in a syntacticised argument structure
L-SUBJECT	*logical-subject* – the semantically most prominent argument
GR-SUBJECT	*grammatical-subject* – syntactically realised subject (pivot)

Figure 2.9: Deconstructing Subject

Arka and Manning (2008) describe AV verbs as those that have the same *a-subject* and *l-subject* realised as the syntactic *gr-subject*. With UV verbs, on the other hand, the *l-subject* is not the same as the *a-subject* in the a-structure. This fact is important in explaining reflexive binding in undergoer voice constructions in Indonesian and justifies the need for this level of representation in the theory (see Manning (1996), and Arka and Manning (2008) for details).

In terms of modelling the processes that the argument structure is subjected to, Alsina (1996:51) notes that the morphological process of passivisation involves an incomplete predicate, and that this 'passive morpheme' with its partial information "must undergo *predicate composition* with a complete predicate to yield a derived predicate" (Alsina 1996:51). The information that the incomplete predicate carries is that the *l-subject* (in Manning's nomenclature) is suppressed.

In a similar way that Alsina (1996) models passivisation as a partial structure combining with a predicate to form a derived predicate, Manning (1996) posits a higher predicate to model constructions such as applicatives, "where the applicative morpheme introduces the higher predicate AFFECT", as shown below, if the higher predicate were to combine with the verb 'separate' (from Manning (1996:44)):

AFFECT < – , – , separate< – , – , – >>

Figure 2.10: The 'higher' predicate AFFECT

Manning (1996) models voice in this way. If we were to represent *locative voice* for Tagalog using this schema, then the a-structure for Example (2.15) would appear as Figure 2.11. It can be seen that the oblique locative adds an extra direct argument in the argument structure.

$$\text{LV} < \underbrace{loc\ ,\ buy< lsubject, theme \mid -}_{} >>$$

Figure 2.11: Representing locative voice in Tagalog

However, with the *undergoer voice*, in Figure 2.12, the *theme* is already a direct argument, and composing the incomplete UV predicate with *buy* does not increase the arity of the derived UV predicate.

$$\text{UV} < \underbrace{theme\ ,\ buy< lsubject, -}_{} >>$$

Figure 2.12: Undergoer voice

The operations initiated by *voice marking* apply a rule that instructs the occupying of the *a-subject* position by a particular constituent with the thematic role licensed by the voice marker. In Indonesian, as we saw in Section 2.2.3, there are only two voice markers, AV and UV. The AV marker allows the logical subject (the most agentive argument to be realised as the grammatical subject) and UV allows the non-agent role, in a transitive construction, to be realised as the grammatical subject. The mechanisms by which these roles are realised syntactically as *subject*, *object* etc, is by a process called *Linking* (see Manning (1996) for details on Linking Theory). Manning (1996) refers to constructions in Tagalog that allow the non-agent role to be realised as the grammatical subject (without demoting the agent) a type of ergative construction. This is much like the UV construction in Indonesian, and for this reason, and for ease of reference, we call this syntactic realisation (where the non-agent is realised as SUBJ, and the agent is realised as OBJ) 'ergative linking', and the AV construction as 'accusative linking'. This 'ergative linking' has also been discussed as a mismatching in prominence in Balinese (Arka 2003:119), where the non-agentive role is mapped onto the SUBJ grammatical.

$$\begin{array}{l} \text{KAN}_{appl} < \text{PRED} < \quad \ldots \quad \overset{\uparrow}{ben} \quad \ldots \quad >> \\ \qquad\qquad\qquad\qquad\qquad\quad \uparrow \\ \qquad\qquad\qquad\qquad\qquad\quad \emptyset \\ \text{I}_{appl} < \text{PRED} < \quad \ldots \quad \overset{\uparrow}{loc} \quad \ldots \quad >> \\ \qquad\qquad\qquad\qquad\qquad\quad \uparrow \\ \qquad\qquad\qquad\qquad\qquad\quad \emptyset \end{array}$$

Figure 2.13: Incomplete predicates for *-i* and *-kan*

The rule for *-kan* applicativisation, as well as *-i*, in Indonesian can also be modelled in the same way, as incomplete predicates, as shown in Figure 2.13. These specify

that an incomplete predicate introduces an applicativised argument. For a the *-i* applicative construction, a locative is introduced as the direct object in the argument structure, and for *-kan*, we have an introduced benefactive object (for details of this applicativised structure see Arka (1993:151)).

Parallelism and Projections Each of the parallel subsystems, or projections, in LFG are governed by their own principles, and are related by projection functions. Figure 2.14 shows how each of the parallel structures – *consituent-structure*, *argument-structure* and *functional-structure* – are related, and the structural correspondences between them.

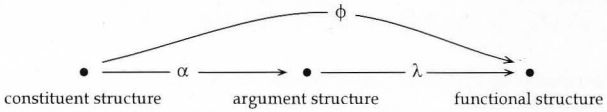


Figure 2.14: Architecture: subsystem of LFG architecture from Asudeh (2004:34)

The connecting arcs represent the functions that relate each of the levels. The c-structure representation describes language in terms of phrase structure trees and the ϕ projection function describes how each node relates attributes and values that represent the f-structure.

So far, we have discussed how linguistic information is represented in the theory, but not how these representations can be generated. The 'instructions' for encoding this information are in the form of annotated rewrite rules, as shown in Figure 2.15

The left-hand symbol (before $\rightarrow$) of each rule expands out to the sequence of right-hand symbols. The annotations serve to constrain the interpretation of the linguistic symbols. These constraining features can also be seen in the f-structure in Figure 2.5.

The previous paragraphs outlined the way in which linguistic information is encoded and related to each other in LFG. In addition, there are principles and conditions that determine the validity of the structures outlined above. These are the well-formedness conditions on the f-structure and principles that guide the construction of the c-structure.

Coherence, Completeness, and Consistency are well-formedness conditions that pertain to the f-structure. Coherence disallows superfluous governable grammatical functions to be present in the f-structure (Dalrymple 2001). This gives us explanatory power for why a sentence such as **The boy sleeps a room* is ungrammatical: as can be seen by the a-structure in Figure 2.5, this construction does not require an OBJ, and it is therefore not licensed. Completeness ensures that the argument list

S	→	NP (↑ SUBJ) = ↓	VP ↑=↓
NP	→	Det	N ↑=↓
VP	→	V ↑=↓	
Det	→	the (↑ DET) =+	
N	→	boy (↑ NUM)=sg	
V	→	sleeps (↑ TNS-ASP PRES)=+ (↑ SUBJ PERS)=3 (↑ SUBJ NUM)=sg	

Figure 2.15: Simple annotated rules required to generate the sentence *The boy sleeps.*

of the PRED is satisfied, and all governable categories are present in the f-structure. For example, simply having the predicate **Sleeps* is not a complete clause. The principle of Consistency disallows incompatible constraints. This principle would disallow incorrect subject-verb agreement, such as **The boys sleeps* resulting in the clashing of number features that are propagated from the verb, and those that are stated in the head of the noun phrase *The boys.*

Finally, a principle that guides the configuration of the c-structure is the **Lexical Integrity Principle** that states that:

> Morphologically complete words are leaves of the c-structure tree and each leaf corresponds to one and only one c-structure node.
>
> (Bresnan 2001:92)

This means that affixes and morphological units do not have positions in the c-structure with the same status as fully inflected lexical items, and that word creation rules and morphological operations belong in the sublexical domain.

2.4 Deviant Lexical Properties

In this section we examine word class categories in the sublexical domain in Indonesian and the morphological processes that operate upon these classes. It has been claimed that certain varieties of Indonesian simply dissolve all open class distinctions (Gil 1994; Gil 2001; Gil 2005; Gil 2010). However, the implementation of the grammar resource we employ (see in Section 3.2.2), and the work we conduct

in deep lexical acquisition rely on the existence of these categories in the language. Therefore, in Chapter 5, we conduct an investigation to verify that these word class distinctions that we had assumed are a linguistic reflection of the language and simply convenient labels for grammar engineering.

In this section, we begin by briefly discussing the nature of word classes and some difficulties on how they are determined, as well as their claimed cross-linguistic significance. The relevant typological claim that we consider is the assertion that certain word classes are universal, and that cross-linguistically nouns and verbs are minimally distinct (Croft 2003). However, when claims are made that a language does not conform to this hypothesis, which has been the case for the dialects Riau Indonesian and Jakartan Indonesian (Gil 1994; Gil 2001; Gil 2010), then the means to verify this claim must be sought, and we report on the previously established methodology and criteria in verifying the justification to dissolve these category boundaries within a language.

In the latter part of this section, we report on the syntactic evidence presented by Yoder (2010), who refutes Gil's (2001) claim. Yoder (2010) shows that the word order facts used as evidence for the dissolving of open class categories can be explained as variations in the canonical word order, such as object or verb fronting, as a focusing strategy, and the unmarked passive construction, which omits the passive prefix *di-* in informal contexts. In a schema laid out by Himmelmann (2008) to explain the mismatch between categorial distinctions in Tagalog, he shows that word classes established within the sublexical domain and the classes of lexical items that occupy the terminal nodes, or *lexical insertion points*, in the syntactic description do not coincide – if word classes are established at the phrase structure level, Himmelmann (2008) claims that it is possible that this distinction is not made at the sublexical level. Yoder's (2010) evidence has tackled one linguistic level, showing that at the lexical insertion points, like Tagalog, Indonesian does discriminate. However, the question remains if Indonesian, like Tagalog, does not distinguish stem categories at the sublexical level. In Chapter 5, we conduct a complementary study showing that Indonesian has word classes, not just from a syntactic point of view as Yoder (2010) proves, but also sublexically.

Word Classes: The Good, the bad, and the ugly

The notion of word classes, such as nouns, verbs and adjectives, is fundamental in both linguistics and computational linguistics. Word classes are the basis for the labels in part-of-speech tagging, and also the building blocks for parsing. In grammar engineering, they are the primitives upon which context-free grammar rules are written. In linguistics, they are considered the categories that shape the organisation of the language. These categories may not align across languages: what is expressed as a verb in one language may be expressed as an adjective or noun in another. But one linguistic universality hypothesis that remains despite these variations is that the

categories noun and verb exist in all languages (Croft 2003).

However straightforward this claim of universality might seem (i.e. minimally, nouns and verbs are distinct), the question of whether these classes are comparable cross-linguistically has been debated. Although certain word classes may be established language-internally, it is more often the case that when comparing different languages, the cluster of features that define the class are not totally coincident, nor would there be total agreement on the members of these classes. The methodology used by linguists in devising a language internal system of word classes is determined in a form-based manner. It is generally agreed upon amongst descriptive and comparative linguists that the determination of the part-of-speech distinctions or word classes *within* a language is determined largely by morphological and syntactic criteria (Schachter 1985; Evans 2000; Haspelmath 2001), rather than the traditional notional categories such as: "a noun is the name of a person, place or thing"; "adjectives denote properties/qualities"; and "verbs denote actions/events", which Croft (2000) points out as having been thrown out without adequate replacement.

The primary criteria in this classification should not be semantic, but based on the grammatical properties of a word (Schachter 1985) . Rather than ascribing to notional parts-of-speech definitions, Schachter (1985) provides these grammatical criteria upon which to base word classes:

- distribution
- range of syntactic functions
- morphological and syntactic categories for which it is specifiable

The method relies ostensibly on the combinatorics of the form (Evans 2000). At the clausal or phrasal level, one looks at how words can combine (or the syntagmatic possibilities of the units) within the phrase or clause. For a structural language like English, we can view this as the position a word can occupy in the c-structure.

When looking at the word level, one investigates how each of the morphological components combine. For example, in English, the suffix *-ly* attaches primarily to adjective stems to form adverbials, such as *slowly* being formed by the affixation of *-ly* to *slow*. One cannot add this suffix to common nouns such as *chair* to form **chairly*. The heuristics for this *combinatorics* approach are outlined by Evans (2000), which takes into consideration semantic or functional properties as a way of labelling these classes rather than determining them.

There are of course problems with taking a thoroughly form-based approach, without any reference to the semantics of the word or stem, or its function, in the determination of word classes, as outlined by Croft (2000), using the English adjective class as a case in point: the adjective *big* in its superlative form is *biggest*, which is combined with the suffix *-est*. However, the adjective *beautiful* cannot combine with the *-est* suffix to form a superlative. Instead, it must be formed analytically with the

adverb *most*, as in *the* ***most beautiful*** *sunset*. Even given this fact about these two stems/words, it may be difficult to argue that these are not of the same word class because in this instance they do not combine in the same way morphologically. One could concede that these form sub-classes of adjectives.

With this form-based approach that is commonly applied in the determination of word classes, Croft (2000) notes that there are two algorithms that can be adhered to in its application. He names these *lumping* and *splitting*. Lumpers strictly apply a morphosyntactic criteria ignoring semantics, they determine a starting point and progressively amalgamate their intermediate word classes through form-based criteria, with tendencies to overload a word classes in a language. On the other hand splitters proceed in the other direction and divide a word class based on minute differences. The major criticism Croft (2000) has with the splitters is that there is no clear stopping criteria. However, ther also seems to be no clear starting criteria either.

Sasse (2001) notes also that these cluster of syntactic rules often lead to 'squishy' categories, and that in many languages there can be a transition from one syntactic word class to another rather than a hard line demarcating the classes; or another kind of word class squish that leads to hybrid categories. For example, Sasse (2001) demonstrates the transition of characteristic properties of *verbs* can stepwise transition into *adjectives* and so on to *nouns*. Such is the case for English, which makes it difficult to determine whether a perfect participle should be classified as a verb or adjective, for example the word *educated* in *He was* ***educated***. Also Malouf (1996) describes the difficulty in analysing gerunds in English given their mix of nominal and verbal properties.

An example of hybrid categories is exemplified in Murrinh-Patha, a language spoken in northern Australia. Based on Walsh's (1996) study, Sasse (2001) reports that the grammatical categories such as noun and verb cannot readily be established in Murrinh-Patha, but instead there are two distinct hybrid morphosyntactic categories called *nerbs* and *vouns*. Both denote qualities, and although *nerbs* are described as more 'nouny' while *vouns* are more 'verby', membership into these categories is based on morphosyntactic criteria such as case inflection, number indicator, and adverb incorporation.

A usually reliable indicator of membership to a word class is inflectional morphology (Haspelmath 2001), however as Musgrave (2001) points out, Indonesian largely has derivational morphology, which is not a criteria usually used in the determination of word classes (see Chapter 5 for more on this topic).

On Deviant Behaviour

Gil's (1994, 2001) main tenet in language analysis is that languages should be described in their own terms rather than being filtered through Eurocentric expectations. He observes that from a syntactic perspective, the labels *noun*, *verb*, and *adjective* in Indonesian are meaningless because effectively all open class parts of

speech can occupy any syntactic slot. Furthermore, Gil (1994, 2005) generalises that the language should only have one syntactic slot *S*. This model of the language reflects that word order is free, and word classes are underspecified.

Croft (2000:67) states that, with respect to claims that a language has no adjectives or there is no noun-verb distinction:

> These assertions are commonly found in reference grammars of languages, most of which are written with no particular theoretical syntactic approach in mind

Because there is no theory that guides the description of the language, the form-based approach that guide how word classes are established are misapplied. Croft's (2000) solution to the misapplication of methodology in the discovery of word classes couches the research of classes within a universal theory of grammar. However, Evans and Osada (2005) take a different approach in righting the misapplication of linguistic methodology, by applying tests on these so-called merged classes for the language in question.

Evans and Osada (2005), who argue against the claim that Mundari, an Austroasiatic language from India, has no noun–verb distinction, have a clear methodology in testing the claim that there are no word class distinctions in a language. They outline that there are three criteria for establishing the lack of word classes within a language. We summarise these criteria Figure 2.16.

i. Equivalent combinatorics
"Members of what are claimed to be merged classes should have identical distributions in terms of both morphological and syntactic categories."

ii. Compositionality
"Any semantic differences between the uses of a putative 'fluid' lexeme in two syntactic positions (say argument and predicate) must be attributable to the function of that position."

iii. Bidirectionality
"[T]o establish that there is just a single word class, it is not enough for Xs to be usable as Ys without modification: it must also be the case that Ys are usable as Xs.

Figure 2.16: Summary of the criteria determining word classes by Evans and Osada (2005)

(2.59) [from Evans and Osada (2005)]

a. *mamuːk=ma* *quːʔas-ʔi*
working-PRS.IND man-DEF
"The man is working."

b. *qu:ʔas=ma* *mamu:k-ʔi*
man-PRS.IND working-DEF
"The working one is a man."

Nootka, a critically endangered language of Canada, is an example given by Evans and Osada (2005) to show the usefulness of their criteria. For the noun *qu:ʔas* "man", and the verb *mamu:k* "work", it would seem that their syntactic and morphological distribution is equivalent. Both are able to act as head of an NP and obtain morphological marking reserved for nominals, and yet also function predicatively and receive aspectual marking, as shown in Example (2.59).

These kinds of examples may lead to the false assumption that the nominal and verbal classes should be merged, as they adhere to Criteria (i) and (ii) in Table 2.16. However, we find that the Bidirectionality criteria fails for certain NP constructions, namely indefinite NP, as seen in Example (2.60).

(2.60) [from Evans and Osada (2005)]

a. *mamu:k=ma* *qu:ʔas*
working-PRS.IND man
"A man is working."

b **qu:ʔas=ma* *mamu:k*
man-PRS.IND working
"A working one is a man."

Although, both the noun and the verb in these examples can equivalently head NPs (as well as VPs), only lexical items from the noun class can head indefinite NPs.

Additionally, Evans and Osada state that these criteria must apply *exhaustively* throughout the language. We address each of the issues in Table 2.16 in our word class investigation in Chapter 5. In this investigation we show how it could be possible to analyse Indonesian as being a language that has no noun-verb distinction, but only if the criteria by Evans and Osada (2005) was partially applied.

Syntactic Reanalysis

Yoder (2010) shows that from a syntactic perspective the free word order exhibited in Riau Indonesian, presented by Gil (1994, 2005) is by and large an effect of discourse focus, and simply variations on the standard SVO word order. Furthermore, he finds that from his quantitative study, word classes are associated with particular functions.

In his study, Yoder (2010) manually examined the 154 examples in Gil's publications, and classified each of the words' lexical categories according to the *Kamus Besar Bahasa Indonesia* (Sugiono 2008). Then for each of the examples, Yoder (2010) classified each word as *default* or *non-default* depending on whether the word's function

	default	non-default
Noun (default=argument)	244 (90%)	27 (10%)
Verb (default=predicate)	160 (99%)	1 (1%)
Adjective (default=modifier)	27 (49%)	28 (51%)

Figure 2.17: Yoder (2010) – Quantitative approach to syntactic classes

i Nominal modifiers (nouns, verbs, adjectives) immediately follow the head noun in the phrase (see Sneddon (1996:142)) – 19 occurrences

ii Equative clauses: The predicate in an equative clause is an NP or AP in juxtaposition with the subject – 22 occurrences

iii Noun phrases can be headless (cf. English: "found reds and a yellow") – 3 occurrences

iv /N-/ rule: The verbal prefix /N-/ changes nouns and adjectives to verbs (see Sneddon (1996:65–66)) – 6 occurrences

v /-kan/ rule: The verbal suffix /-kan/ changes adjectives to verbs – 2 occurrences

Figure 2.18: Yoder (2010) – Accounting for Lexical Exceptions

coincided with the expected word class, for example, if a noun was found to function as an argument or a verb as predicate, then this would be classified as *default*. The statistics he discovered are shown in Figure 2.17. These show that there is a strong correlation between form and function.

Furthermore, the exceptions he found (*non-default*), can be explained by grammatical rules that are imposed on Standard Indonesian, which are listed in Figure 2.18

Both items (iv) and (v) in Figure 2.18 pertain to lexical rules that change the word classes,[18] while the items (i) to (iii) pertain to syntactic configurations found in Standard Indonesian.

Yoder (2010) uses the same 154 sentences to establish that the default word order in this dialect of Indonesian is S V O, with 76% of all utterances following this order. Of the other 24%, these variations can be explained through various discourse functions. The remaining 24% of so-called free word order, are accounted for through: (1) object fronting for focus; (2) verb fronting for focus; (3) subject "after-thought" post-posing; and (4) an unmarked passive construction.

For Tagalog, a language also claimed to have no distinction between words in

[18]The prefix /N-/ in (iv) is a variant of the *meN-* prefix in Standard Indonesian.

open class categories (Kaufman 2009), these functional characteristics Yoder (2010) describes for Indonesian seem to be in parallel – Himmelmann (2008) shows that at the syntactic level, there are also these functional slots Yoder (2010) shows for Indonesian. There are words licensed to fill certain syntactic positions. Even though these syntactic slots that coincide with their function exist, at the sublexical level all stems may undergo any morphological process. That means that, in Tagalog, there is a difference between morphologically complete words that occupy syntactic slots and stems that undergo morphological processes at the sublexical level. In Chapter 5, we test to see if Indonesian, like Tagalog, dissolves the differences between open class categories sublexically, while maintaining word class specific functional positions at the syntactic level for morphologically complete words.

Final Remarks

Although it is generally agreed upon that a cluster of morphological and syntactic rules form the basis of determining word classes, rather than notional categories (Haspelmath 2001; Schachter 1985), it has been shown that this method does not always produce well behaved classes, for example *vouns* and *nerbs* in Murrinh-Patha. In addition, there are ways of applying this form-based criteria, by *lumping* or *splitting*, which contribute their own difficulties, and does not steer the linguist away from misanalysis, particularly in having a heavy-handed approach in lumping that leads to an analysis that dissolved word class distinctions. However, in such cases Evans and Osada (2005) formulate criteria to apply to controversially merged categories.

2.5 Lexical Semantics as a Determinant of Variation in Argument Structure

The previous section discussed the difficulty in defining word classes language-internally, which poses challenges in cross-linguistic comparison. In this section, we briefly discuss the hypothesis that verbs, given they can be established in a language, can be grouped into subclasses according to *diathesis alternations*, based on lexical semantic similarity. Levin (1993) hypothesises a tight connection between a verb's syntactic behaviour and its meaning, and produced an extensive evidence in English in support of this claim.

Diathesis alternations describe the range of possible combination of arguments verbs can take, which accompany a slight change in meaning for each alternation. However, not all verbs can participate in the same range of alternations, which forms the criteria for the subgrouping of verbs. For example there are alternations that allow an intransitive expression of a verb, as shown with *break* in Example (2.61).

(2.61) [from Levin (1993)]

a. The window broke. (inchoative)

b. The little boy broke the window. (causative)

Levin (1993) labels this pair of alternations the *inchoative/causative alternation*, which the verb *break* can participate in. Although, *appear* can occur in the inchoative construction, it cannot head a causative construction in the same way, as shown in Example (2.62).

(2.62) [from Levin (1993) optional bracketing added to mirror Example (2.61)]

a. The rabbit appeared (out of the magician's hat).

b. *The magician appeared a rabbit (out of his hat).

Verbs that participate in this inchoative/causative alternation, and therefore are similar to *break*, are *shatter*, *smash*, *tear*, among others (Levin 1993:28). The verbs that are grouped with *appear*, and are unlikely to not participate in the causative alternation, are *emerge*, *erupt*, *flow*, *materialise*, to name a few (Levin 1993:258).

The assumption that the syntactic behaviour of verbs is semantically determined has been widely investigated in Computational Linguistics, particularly for *deep lexical acquisition* as we discuss in the following section. However, as we will see the assumptions made with these studies shows that syntactic structure is a good determinant of semantic similarity. To our knowledge, very few studies utilise semantic similarity to infer syntactic features. One such study by Baldwin (2005) uses WordNet[19] to construct a set of semantically similar words for a particular sense of a word. In this study, Baldwin (2005) aims to produce a lexical entry for each target word, specifying syntactic type[20], by taking a majority vote of the syntactic types of all the words deemed to be semantically similar. The results showed that at the type level, their method of using synonyms to predict syntactic information did not exceed the baseline.

2.6 Deep Lexical Acquisition

The manual creation of lexicons is an expensive endeavour in the production of a precision grammar, but it is a fundamental component. Baldwin *et al.* (2005) discovered that the leading cause of parse failures in an experiment conducted with the English Resource Grammar (Copestake and Flickinger 2000) was due to missing

[19]A lexical database of nouns, verbs, adjectives and adverbs grouped into a set synonyms, called *synsets*. These groups of synsets are related to each other via a structure that encodes their super-subordinate relation (`http://wordnet.princeton.edu`).

[20]These syntactic types are fine-grained syntactic categories which can have information about the kinds of dependencies allowable for that word

lexical entries, which were either out-of-vocabulary items or missing subcategorisation frames for an existing verb. This contributed to 41% of parse failures, and ranked second to this was constructional gaps (i.e. failure due to a syntactic rule not being encoded). In addition, Briscoe and Carroll (1993) found that half of the failures in their experiments on parsing unseen data were due to lack of lexical entries with appropriate subcategorisation frames. Within the ParGram community, efforts have been made to mitigate the problem of having incomplete or partial lexical information. Crouch and King (2005) had designed a method of creating a lexicon that could merge information from a variety of lexical resources to identify and fill missing syntactic information.

Baldwin (2005, 2007) defines the area of Deep Lexical Acquisition (DLA), in which a given lexeme is mapped onto a system of predefined lexical types. DLA methods enable the automatic acquisition of detailed lexical information that is required for a deep grammar, such as verb subcategorisation acquisition, the classification of nominal classifiers, distinguishing adjectives that may only be used predicatively, and classifying nominals as count or mass nouns.

DLA tasks have been successfully performed on Malay, a language related to Indonesian with 80% lexical overlaps (Gordon 2005). For example, Nicholson and Baldwin (2008, 2009) successfully employ a maximum entropy learner to learn count classifier preferences for nominals in Malay. However, this study pertains to the automatic classification of nominal information, and the research we conduct in DLA is associated with verbal information. In this domain we discuss the methods used in acquiring subcategorisation information or the discovery of verbs that share the same syntactic profiles in the same style as the Levin (1993) classification. However, in order to conduct DLA on the verbal class in Indonesian, this class should first be established, given the controversy surrounding their existence (Gil 1994; Gil 2001; Gil 2010). The experiments in verifying word classes has a two-fold purpose to: (1) ensure that the verbal class is distinct from other open class lexical items in order to perform DLA; and (2) to verify the existing implementation of the grammar resource described in Section 3.2.2 and Chapter 4, which assumes the existence of the verb class as separate from other classes. Later, in this section, we also outline some of the methods used in the induction of part-of-speech information from a corpus, as a means of determing whether there are word class distinctions (within the open class category) in Indonesian.

2.6.1 Acquiring Verbal Information

With respect to the acquisition of lexical information from corpora, much of the work conducted in this field requires the use of NLP tools and resources not available in Indonesian. For example, studies such as O'Donovan *et al.* (2005) rely on the structural information already encoded in the syntactically labelled Penn Treebank in order to automatically annotate grammatical information in the f-structure of a

deep LFG grammar. On the other hand Brent (1993) composes a series of algorithms to extract subcategorisation information from *Brown Corpus* and employs a system based on deterministic morphological cues to identify predefined syntactic patterns to map verbs onto.

With Indonesian being a relatively under-resourced language for NLP (see Section 2.2.1), many tools and corpora available in other languages are not available to us in our investigation. Therefore our approach is to discover semantically similar classes of verbs that behave in the same way syntactically, exploiting Levin's (1993) hypothesis. In the discovery of Levin classes, or the grouping of verbs according to *diathesis alternations*, the approaches in achieving these tasks minimally requires three fundamental components: (1) resources; (2) acquisition method; and (3) evaluation (Korhonen 2010; McCarthy 2006; Schulte im Walde 2009).

Resources

The resources used in these tasks rely in part on the type of acquisition method that is used. The data employed can be labelled or unlabelled.

Labelled data are encoded by human experts, or even expert systems, the information that we want to automatically learn (*gold standard data*), or information that provides annotated linguistic features that are helpful for the task (*tagged corpora*). For example, if our task is to learn whether a verb is transitive or intransitive, then our gold standard data may be a list of verbs that specify transitivity, and we may employ a parsed corpus, such as the Penn Treebank to assist in building a system that predicts transitivity.

Examples of labelled data are corpora such as the Penn Treebank (Marcus *et al.* 1993), which has information about syntactic structure for portions of the *Wall Street Journal*, and the Brown Corpus (Francis and Kučera 1979), which has added part-of-speech information to a variety of printed genres including press reportage, theatre reviews, religious text, letters, and biographies.

Any collection of text can be used as an untagged corpus. However, there are many NLP techniques that can be employed to process the data to acquire more linguistic information. For example, Manning (1993) employed a shallow parser (chunker) in order to acquire subcategorisation information from the *New York Times*. Schulte im Walde (2006) induced subcategorisation information for German with the use of a lexicalised probabilistic context free grammar (PCFG), and Joanis *et al.* (2008) achieved this with the use of a part-of-speech tagger and chunker. These are examples of ways that relevant features are extracted from raw text to guide learning, in the acquisition of lexical information.

While many lexical acquisition tasks employ resources that pertain to the use of syntactic information (Manning 1993; Schulte im Walde 2006; Sun *et al.* 2010; Sun and Korhonen 2011), not all studies rely on this feature but instead rely on collocations (Li and Brew 2008). Unfortunately many of the studies that utlise lin-

guistic data prediction, such as predicting syntactic categories, in the quest to further study verb classes focus mainly on English (Merlo and Stevenson 2001; Parisien and Stevenson 2010). This is because, as Täckström *et al.* (2012) state, it is the language with the most resources. Some methods that have been used to mitigate the problem of addressing the lack of annotated resources, or the cost of manual annotation are unsupervised and semi-supervised (Klein and Manning 2004; Koo *et al.* 2008; Søgaard 2012), and transfer learning methods where poorly resourced languages leverage the data developed for well-resourced languages (McDonald *et al.* 2011s; Täckström *et al.* 2012; Naseem *et al.* 2012). Although it does not entirely eliminate the need for any annotated data, unsupervised and semi-supervised methods can still benefit from the vast amount of unannotated corpora or plain text available ripe for unsupervised learning.

One final method of automatically generating linguistic annotation for under-resourced languages is simply collecting them through targetted publicly available sources on the Internet (Xia and Lewis 2009). Unlike the aforementioned methods, this method gathers data that has already been constructed by linguists,[21] rather than using machine learning methods to automatically generate more annotation. One advantage to this method is that the annotations are more reliable, having been produced by experts rather than generated (semi-)automatically. The disadvantage is that there is no guarantee that the format or the kind of annotation available is what is required for the task at hand. Also, the amount of annotated data is far more limited than the amount of plain text available.

Acquisition Method

Although this dichotomy is not absolute, there are two broadly defined techniques applied in lexical acquisition, which are *supervised* (*classification*) and *unsupervised* (*clustering*) methods.

In general, unsupervised methods learn from unlabelled data, or does not make use of labels from a human expert to guide learning. As noted by Smith (2011:110), when scientists first encounter the notion of unsupervised learning, they find it difficult to rationalise the learning of linguistic structure from data that has not had desired information added by experts:

> '...how could it be possible to learn to predict an outcome without ever having seen one?' It is important to remember that unsupervised learning is not "free." It requires some understanding of the domain, and the development of an algorithm that incorporates prior knowledge...

The means by which we can incorporate knowledge to guide learning is through the features we develop, and through linguistic assumptions we make to infer meaningful associations, as in Levin's assumption that syntactic alternations are closely

[21] in the form of interlinear glossed text

Data Set	# Verbs	# Classes
GS1	835	15
GS2	204	17
German-SIW	168	43
French-Sun	116	16

Figure 2.19: Gold Standard Data Sets

tied to lexical semantics. Unsupervised learning entails the use of clustering methods. Clustering aims to gather items of interest into groups that share some likeness, and separate them from items that are unlike them. Schulte im Walde (2006) employs Hierarchical Agglomerative Clustering (HAC – see Manning *et al.* (2008:381)), and compares this with K-Means (see Manning and Schütze (2000:515)), in her experiments to automatically induce Levin-style German verb classes. The advantages of hierarchical clustering over flat clustering (such as K-Means) is that with hierarchical clustering algorithms, there is no need to prespecify the cardinality of the clusters (the number of clusters we expect the algorithm to generate). However, in her experiments, Schulte im Walde (2006) uses HACE to initialise her clusters, and further processes them.

Supervised methods employs the gold standard data to guide learning. For example, Merlo and Stevenson (2001) classify verbs as belonging to three optionally intransitive classes according to their diathesis alternations, namely unergative, unaccusative, and object-drop verbs in English. Based on a collection of linguistic features, which they infer from an automatically part-of-speech tagged and parsed *Wall Street Journal* corpus, they employ the updated version of the C4.5 algorithm, C5.0, an iterative data mining algorithm used to produce a decision tree (see Witten and Frank (2005:169) for details on C4.5 and decision trees as classifiers). A split is produced at each node in the tree based on a linguistic feature that provides the highest information gain value, with respect to the separation of verbs according to their class.

In this thesis we apply unsupervised methods, which we discuss in more detail in Section 3.3.

Evaluation

Korhonen (2010) reports that there are two gold standard datasets (GS1 and GS2) used in many of the studies that automatically label verb classes according to Levin's categorisation in English. These English gold standard datasets are summarised in Figure 2.19: GS1 comprises 835 verbs in 15 broad and fine-grained classes (Joanis *et al.* 2008), and GS2 has 204 medium-to-high frequency verbs in 17 fine-grained

Levin classes (Sun *et al.* 2008).

Each class in the English gold standard Levin-style datasets GS1 (Joanis *et al.* 2008), and GS2 (Sun *et al.* 2008) are not as sparsely populated as the non-English data; German-SIW (Schulte im Walde 2006) has 168 verbs distributed into 43 broad and fine-grained classes, while French-Sun (Sun *et al.* 2010) has 116 French verbs populating 16 broad and fine-grained Levin classes.

In the development of German-SIW, verbs were manually classified into their Levin-style classes based mainly on intuition, with a close mirroring to Levin's (1993) English classes. The reliance on an already established English gold standard and the use of reference material, such as bilingual and monolingual dictionaries, and corpus searches for verification in order to produce the classes, were the reasons why inter-annotator agreement was not performed in the creation of this dataset (Schulte im Walde 2006:161-164)

Likewise, French-Sun, the French dataset (Sun *et al.* 2010) was manually created using the French translations of Levin's (1993) English classes. In addition, Sun *et al.* (2010) explicitly consider diathesis alternations of each of the candidate verbs, and not just lexical semantics to obtain the gold class. They omitted verbs from certain classes if they did not adhere to the same syntactic alternations as the other members of the class.

For the tasks undertaken by Sun *et al.* (2008), Schulte im Walde (2006), and Sun *et al.* (2010), the evaluation measures employed in appraising the goodness of the classes induced or classification applied to verbs in these tasks are variations on F-score. It is a metric that takes into account the number of verbs that are correctly labelled as belonging to a particular class, and balances it with the number of verbs that rightly belonged in that very class but were incorrectly labelled otherwise. Although all three of these studies employ this metric, the way in which the measure is calculated varies somewhat. For example Schulte im Walde (2006) employs the paired F-score, which evaluates each verb discovered in an induced cluster pairwise, while Sun *et al.* (2008), and Sun *et al.* (2010) do not (see Section 3.3.6 for details on evaluation metrics). On the other hand, Joanis *et al.* (2008) do not employ a variation of the F-score, but calculate the soundness of their method with a measure called *accuracy* (see Manning *et al.* (2008:115) for details).

The GS1 and GS2 have become de facto English datasets for the acquisition of Levin-style classes for English. Korhonen (2010) summarises the state-of-the-art systems for Levin-style systems for English verbs. The top systems achieve in the ball park of 66% accuracy or 80% F-score, for GS1, and GS2, respectively. For French and German, although the same metrics are employed, the types of datasets that are used in evaluating these tasks are quite different, and pitting the English results against the French and German results may not be an accurate representation of performance. To begin with German-SIW and French-Sun are less densely populated per class. Schulte im Walde (2006) achieves a score of 22.19%, which seems much lower than the English state-of-the art. But when applied to unknown verbs, the performance of

the system almost halves. However, Schulte Im Walde's method, which clusters verbs based syntactic features derived from a probabilistic context-free grammar (PCFG), far surpasses the random baseline.

Sun *et al.*'s (2010) French system, which groups verbs according to VerbNet, taking into account the verb's thematic grid, achieves its highest F-score of 55.1% using concurrence features, which performed better than their subcategorisation features. Also, Falk *et al.* (2012) partially map their verb classes onto the French-Sun dataset in order to better gauge the performance of their system. Although the mapped gold standard dataset is not exactly the same, they do report a 70% F-score on a similar data set, employing a neural clustering method, using a mixture of subcategorisation, syntactic and semantic features.

In terms of methodology, the studies that we look to are those systems that are built to disambiguate and discover syntactico-semantic Levin-style classes, rather than systems that aim to induce valency or syntactic frame information from corpora, such a O'Donovan *et al.* (2005) and Manning:1993, respectively. Lexical acquisition systems can be built in a supervised fashion as in Lapata and Brew (2004) or tackled as a clustering task as in Schulte im Walde (2006) or Bonial *et al.* (2011). Lapata and Brew (2004) develop a semi-supervised system that generates, for a given verb and its syntactic frame, a probability distribution over the Levin verb classes. They then use this system to disambiguate tokens using collocation information. Our system, like Schulte im Walde (2006) uses an unsupervised clustering approach. In her approach, Schulte im Walde (2006) employs hierarchical agglomerative clustering over parse features to discover word classes in German, and evaluates using manually created gold standard data.

2.6.2 Part-of-Speech Induction

Part-of-speech induction aims to approximate syntactic labels based on unlabelled token sequences, usually resulting in the discovery of categories that are not considered linguistically motivated (Biemann 2009). It is traditionally approached using unsupervised methods, with the assigning of labels being done after the fact. However in our study we use linguistic features in order to induce linguistically motivated clusters. Given that the linguistic methodology employed in determining word classes is heavily form-based (see Section 2.4), the unsupervised methods employed in part-of-speech induction, in conjunction with the linguistic features we employ, are well-suited for our word classes experiments to see if verbs are indeed distinct from nouns.

Features There are two main feature types employed in part-of-speech induction systems: morphological, and collocational (local syntactic context). Christodoulopoulos *et al.* (2010) conducted a survey of 7 part-of-speech induction systems, 5 of which developed systems based on solely on collocational or distributional properties, while

two systems (Clark 2003; Berg-Kirkpatrick *et al.* 2010) additionally included morphological features. Christodoulopoulos performed system comparisons to determine which methods were best suited for the task. The systems were trained and evaluated on the full *Wall Street Journal* portion of the Penn Treebank. The best performing systems from this survey paper were those that employed morphological features along with context features. The system by Clark (2003) was very similar to another in the survey, by Brown *et al.* (1992). However Brown *et al.*'s (1992) system did not incorporate any morphological features in their system, and Christodoulopoulos *et al.* (2010) attribute Clark's (2003) superior performance over Brown *et al.* (1992) to the use of morphology.

Methods All part-of-speech induction systems employ unsupervised methods. Berg-Kirkpatrick *et al.* (2010) take a completely unsupervised approach to part-of-speech induction, where they only make use of the unlabeled text itself. They optimise using the Expectation Maximisation algorithm over their locally normalised hidden Markov models (HMMs). Clark (2003:62) uses a similar model, with his models giving a higher probability to partitions that have words with similar morphologically strings in the same cluster.

Relevant to the morphological features we develop in our study is the research of Goldsmith (2001), and his use of the term *signatures* to mean a collection of morphological patterns relevant to a stem. He employs the MDL (minimum description length) algorithm as a way of discovering and grouping verbs that inflect in the same way in English (or allow the same pattern of affixes to the stem). Although most work in part-of-speech induction employ collocational or distributional features that emulate syntactic position, Goldsmith's work focuses solely on morphological patterns as a way of grouping *inflectional categories*. Rather than inducing the signatures in the way the Goldsmith does, we derive them from a corpus using a non-class biased morphological analyser (See Section 5.4 for more details).

Chapter 3

Tools and Resources

This chapter outlines the tools and resources we employ throughout this thesis. In Section 3.1, we briefly present the grammar engineering platform used in implementing the linguistic analyses in our case studies.

In Section 3.2, we describe the grammar resources we employ, and build upon. Many of the tools and resources that we employ in this study are those that are used within the **ParGram** community, such as XLE and XFST discussed in Section 3.1. Research groups within the ParGram project aim to develop grammars and resources in parallel, such as the ParGramBank treebank (Sulgar *et al.* 2013), as discussed in Section 3.2.1. The grammar we use within this study, **IndoGram**, is part of this parallel development project, which we introduce in Section 3.2.2.

3.1 Grammar Engineering Tools

3.1.1 XLE: Grammar Development Platform and Parser

XLE is a platform for developing, testing, and debugging large-scale deep grammars that are encoded according to the theory and principles of Lexical Functional Grammar. It consists of efficient algorithms for parsing and generating, and a transfer engine, which is used in tasks such as generation and machine translation. It also provides a graphical development environment, which enables the user to inspect the output of a grammar in a convenient way for debugging.

The Parser

The XLE parser is an active state chart parser, especially optimised for unification grammars such as LFG. Such optimisations are necessary because given the required computation to determine whether an input string is valid for the given grammatical description, in the worst case, the parse time can be exponential to the length of the input (Maxwell III and Kaplan 1994; Maxwell III 2012).

Maxwell III (2012) describes three key ideas in making the XLE parser efficient, which are:

1. Exploiting the context-free nature of the phrasal contraints in order to minimize the time impact of computing the functional constraints;
2. Employing contexted unification, which is an algorithm for merging alternative or disjunctive feature structures together;
3. Applying an optimisation technique called 'lazy contexted copying' during unification, which involves the marrying of two techniques during unification.

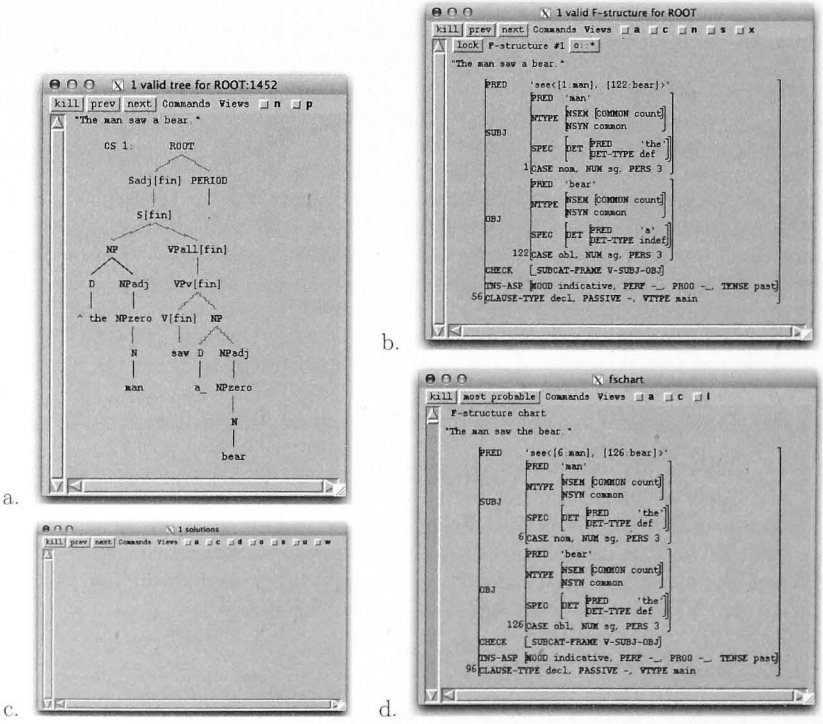


Figure 3.1: Window for inspecting feature structures for corresponding c-structure.

The Grammar Development Environment

XLE provides a platform for grammar development, testing, and debugging, which are accessed via a Tcl shell interface. This interface uses Tcl syntax to execute commands available in XLE, such as loading grammars, parsing sentences, or running tests.

There is a comprehensive manual for XLE (Crouch *et al.* 2011); the *XLE User Documentation* gets the new user started with installation, and familiarising the user with the platform to the stage where they can begin entering a grammar, and beyond. There is also a starter's guide, the *Walkthrough*, for those new to the development platform or those who are familiar with LFG but not grammar engineering. The Walkthrough guides the reader through the process of entering a new grammar, from introducing the basic components (and files required) to make a up a working grammar to creating rules, and lexical entries, and debugging and testing.

There are a number of ways to inspect and export the analyses generated from the grammar via the Tcl graphical interface. Parsing a single sentence or phrase via the shell interface initially generates four windows: the *tree*, *f-structure*, *solutions*, and *packed f-structure* windows. The *tree* and *f-structure* windows reveal the c- and f-structure possibilities, as generated by the grammar. The *fs-chart* window displays the packed f-structure which presents all the possible f-structure analyses in the one chart. The *solutions* window presents partial f-structures, which compares directly the alternative attributes and values that is generated by the grammar.

Parsing a sentence, such as *The man saw a bear* produces the four windows shown in Figure 4.5. For an unambiguous sentence such as this, the (b) f-structure and (d) fs-chart are identical. The *solutions* window, shown in (c), which presents the user with f-structure fragments, with possible f-structure alternatives is empty.

Morphological analyses can also be inspected via the graphical tool, for the solution at hand, or all possible analyses for a given input string via the commandline shell interface. The latter simply reveals the output of the morphological analyser, while the former presents a solution in context. The possible analyses of strings offered by XLE are simply the output of the morphological analyser, which we introduce in Section 3.1.2.

3.1.2 XFST: Finite State Tools

XLE also integrates finite state tools for morphological analysis (Kaplan *et al.* 2004). It provides an easy interface to finite-state calculus algorithms, in particular the Xerox Finite-State Calculus implementation (Beesley and Karttunen 2003). The finite-state network we create with these tools is a transducer, which allows for a 'lower language' — or a definition of the allowable surface words in the language — and an 'upper language;, which defines the linear representation of the morphological units in the surface word, as shown in Figure 3.1.2.

surface form (upper language)
↑
intermediate forms
↓
analysed form (lower language)

XFST allows for the encoding of concatenative morphology, but can also accommodate non-concatenative morphology such as reduplication via a function called 'compile-replace' (Beesley and Karttunen 2003). This function is applied to substrings between parenthetical tags. Also relationships between affixes that are interdependent, but are non-adjacent, can be signalled via a built-in mechanism called *flags*. These flags enable the enoding of circumfixes in Indonesian, such as the nominaaliser *pe... an*.

3.2 Grammar Engineering Resources

3.2.1 ParGram

ParGram is a consortium of researchers and research groups who aim to develop large-scale parallel grammars. The parallel grammar development effort started with three languages: French, English, and German (Butt *et al.* 1999b; Butt *et al.* 1999a; Rosén and Zaenen 1999), and has now expanded to more than a dozen languages, including Urdu, Turkish, Japanese, and Indonesan, each at various stages and rates of development.

The commonalities of the grammar (as well as their deviations from each other) are represented in the f-structure as identical attribute labels. The manner of usage and interpretation of overlapping attributes and values employed in the parallel grammars are an agreed-upon set by the ParGram community. Unfortunately, there is not a complete and up-to-date repository of this inventory of agreed upon ParGram features and grammatical functions. However, the Starter Grammar[1] lists some of the more established features in use, and Butt *et al.* (1999b) exhibit some of the established ways of encoding particular grammatical constructions.

The need for this rigourous standardising of implementation and nomenclature is imperative, because even in linguistic studies for different language families the same term may mean different things and be applied in different ways, however slight or substantial. For example the term **focus**, among Austronesianists, used to refer to

[1] `http://www2.parc.com/isl/groups/nltt/xle/doc/PargramStarterGrammar/starternotes.html`

the noun phrase that the verb picked out as the most prominent entity in the clause by way of an affix (see Blust (2002) for discussion on this terminology). However, for non-Austronesianists, the term focus is synonymous with *rheme*, which is the new information in the clause. In general, Austronesianists now use the term *voice* for this verbal phenomenon (see Section 2.2.3 for more on voice).

There are many benefits to developing grammars in a parallel manner. From a linguistic perspective, it provides researchers a way to seek out the commonalities in languages, and possible variation and to formalise them in a systematic way. Also, given that the interpretation of the attributes and values are agreed upon, it allows for greater transparency in a transfer system for machine translation.

In addition to a framework for linguistic inquiry and the tools to conduct such investigations, ParGram provides a library of resources for grammar development. One such resource is the Starter Grammar, which serves as a template and a guide for building a new grammar and provides a skeletal architecture.

The Starter Grammar is an example of a baby English LFG grammar that employs the annotation schema that can be interpreted by the XLE parser. LFG annotations, such as the $\uparrow$ and $\downarrow$, are replaced with alternatives '^' and '!', respectively. The full set of corresponding LFG-to-XLE annotations can be found in (Crouch *et al.* 2011). Throughout this thesis we also employ the symbols parseable by XLE when referring to rules written in the Indonesian ParGram (IndoGram) grammar, but also add the traditional LFG operators when the symbols are not equivalent.

3.2.2 IndoGram

IndoGram was developed at The Australian National University with the aid of the ParGram community.[2] It has 106 rules in total encoding both the syntactic rules and word formation (sublexical rules) employing the XLE schema. There are approximately 2000 common nouns with English glosses, of which almost one quarter are multiword expressions, and 150 verbs with the required syntactic information for parsing.

There are two separate but interrelated lexicons: a morphological lexicon, using the syntax required by *lexc*. *Lexc* is a tool used with the XFST suite (Beesley and Karttunen 2003) that we use for storing the stem lexicon with their appropriate word class information for word formation in the morphological analyser. The XLE lexicon has information that is used by the parser and has more detailed syntactic information, such as subcategorisation frames.

The rules in the IndoGram resource was developed using the ParGram parallel testsuites. These tests developed by the ParGram community (see Sulgar *et al.* (2013) for the kind of linguistic phenomena and diversity of constructions covered in the design of the testsuites). We repeat the testsuites Appendix A as a means to show

[2]http://pargram.b.uib.no/research-groups

the coverage of the IndoGram grammar.

Each testsuite has its own theme, for example *The Fall 2009* sentences focus on nominals and nominal phrases, particularly on the type of lexical head that is exhibited. Here, lexicalised noun compounds were entered in the lexicon as multiword expressions (MWE),[3] but compositional MWEs are parsed as noun compounds in the grammar.

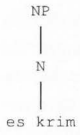
An example of a MWE is Example (3.1) *es-krim* "ice cream", and *kulit buku* "book cover" is parsed as a noun compound (NN), as shown in Example (3.2).

(3.1)

"es krim"

```
[PRED  'es krim'                          ]
[CHECK [_STEM noun]                       ]
[NTYPE [NSEM [COMMON count]]              ]
[      [NSYN common       ]              ]
1[PERS 3, TRANS ice-cream                 ]
```

CS 1:

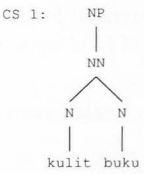


(3.2)

"kulit buku"

```
[PRED  'kulit'                                                          ]
[      {[PRED  'buku'                                                   ]}
[MOD   {[CHECK [_STEM noun]                                             ]}
[      {[NTYPE [NSEM [COMMON count]]                                    ]}
[      {[      [NSYN common       ]                                     ]}
16[MOD-TYPE compound, PERS 3, TRANS book                                ]
[CHECK [_STEM noun]                                                     ]
[NTYPE [NSEM [COMMON count]]                                            ]
[      [NSYN common       ]                                            ]
3[PERS 3, TRANS skin                                                    ]
```

CS 1:



IndoGram's Morphological Analyser is based on Pisceldo *et al.* (2008) and further developed by Mistica *et al.* (2009), and encoded using the XFST tools described in Section 3.1.2. At present, there is a mismatch with the two lexicons in IndoGram:

[3]Non-lexicalised noun compounds are parsed. We assume these exhibit *productive compounding*. The term 'productive compounding' can be gauged on a sliding scale making it difficult to determine what actually should be considered a MWE that is indeed lexicalised. One criteria for determining this is by being able to identify the semantic relationship between the nominals in the noun compound that occur regularly and predictably. Girju *et al.* (2009) have identified a set of 7 semantic relations of this kind through empirical corpus investigation:

SEMANTIC RELATION	EXAMPLE
CAUSE-EFFECT	*laugh wrinkles*
INSTRUMENT-AGENCY	*laser printer*
PRODUCT-PRODUCER	*honey bee*
ORIGIN-ENTITY	*alien message*
THEME-TOOL	*news conference*
PART-WHOLE	*car door*
CONTENT-CONTAINER	*apple basket*

there is a large stem lexicon for the morphological analyser and a much smaller XLE lexicon. The stem lexicon for the morphological analyser is much larger because it only contains part-of-speech information, that is based largely on the description in the KBBI *Kamus Besar Bahasa Indonesian* "The big Indonesian dictionary" (Sugiono 2008).

3.3 Natural Language Processing Tools and Resources

3.3.1 Data

The data we use in our investigations for Chapters 5 and 6 use articles from Wikipedia. In particular we used a dump of a snapshot of the Indonesian Wikipedia[4] because not only is it a large source of text, but also because the data is produced and curated by many authors; it is representative of the way the language is used throughout the Internet-connected areas of Indonesia, and Indonesian speakers throughout the world. For this reason we chose this 'crowd-sourced' open encyclopaedia as our text collection.

We gathered approximately 26 million Indonesian tokens from Wikipedia articles and removed the mark-up used by Wikipedia for hyperlinks and other mark-up used in rendering the article. In the preparation of the Wikipedia data, we use WikiPrep[5] to remove the mark-up. We ran a sentence and word tokeniser over the text using two tools for comparison, namely TokLem[6] and OpenNLP.[7] We performed sentence tokenisation as a first step mainly as a preparation for word tokenisation because this process often distinguishes between abbreviations from sentences final punctuation.

Neither tool is designed specifically for processing Indonesian, however we test their performance using a small sample from the Indonesian Wikipedia dump. We took 453 sentences with 9,139 tokens and hand analysed them for evaluation. TokLem is an all-in-one sentence detector and tokeniser designed specifically as a test case for processing Malay, and is evaluated on very small corpus of Malay online newspaper articles (Baldwin and Awab 2006). It was built in the flex environment based on hand-coded rules.

We also tested the Maximum Entropy (MaxEnt) models trained on a comparatively large English dataset built from the OpenNLP tools,[8].

Table 3.1 shows the F-score for Toklem and OpenNLP on the 453 Wikipedia

[4]We used the dump produced on October 15, 2009, downloaded on October 19, 2009 from `http://dumps.wikimedia.org/idwiki/`

[5]`http://www.cs.technion.ac.il/~gabr/resources/code/wikiprep`

[6]`http://code.google.com/p/malay-toklem/` accessed 14/01/2010)

[7]`http://opennlp.sourceforge.net/projects.html` accessed 14/01/2010

[8]v1.4 available from `http://opennlp.sourceforge.net/models-1.4/`

[t] System	SENTENCES	TOKENS
TokLem	0.890	0.969
OpenNLP	0.956	0.991
Total Number	453	9139

Table 3.1: Tokenisation for Wikipedia

sample sentences. Based on these results we used OpenNLP for sentence and word tokenisaion for the rest of the Wikipedia data.

We also use Wikipedia, not only in our experiments in the latter part of this dissertation, but also as text source Section 4.4, where we map out clusters of *-kan* alternations in extending the Indonesian XLE lexicon.

3.3.2 WEKA Toolkit

We use an off-the-shelf tool WEKA for our soft-clustering in Chapter 5 for our word classes investigation. In particular we use the EM implementation. Soft clustering allows a probabilistic membership into classes, unlike K-Means. Manning and Schütze (2000) describe the EM algorithm as a 'soft' version of K-Means.

K-Means is a hard clustering method, which means an item or instance is assigned to exactly one cluster. For initialisation, K number of centroids are first determined, and iteratively each instance is assigned to a cluster whose centroid is closest.

EM-Algorithm The first component of the EM algorithm (*Estimation Step*) is the same as the process by which K-means assigns a data point to a cluster, however at the second step of EM (*Maximisation Step*), the centroids are recomputed such that the likelihood of the parameters of the distributions are maximised. The estimation and maximisation steps are repeated iteratively until the parameters do not change or reach a specified threshold (Tan *et al.* 2006).

3.3.3 *hcluster*

For our hierarichal clustering we used *hcluster* a Python scripting language add-on.

Hierarchical Agglomerative Clustering HAC is a bottom-up clustering algorithm summarised by Jain *et al.* (1999:p277) in these three steps:

1. Compute the proximity matrix containing the distance between each pair of patterns. Treat each pattern as a cluster.

2. Find the most similar pair of clusters using the proximity matrix. Merge these two clusters into one cluster. Update the proximity matrix to reflect this merge operation.

3. If all patterns are in one cluster stop. Otherwise, goto step 2.

hcluster has three linkage criteria for merging clusters in step 2: (1) complete linkage clustering; (2) weighted linkage clustering (WPGMA); and (3) average linkage clustering (UPGMA) (see Manning *et al.* (2008:381) for more on linkage criteria).

In complete linkage clustering, merging is determined based on the furthest pair of instances within the cluster, while in WPGMA and UPGMA, the decision to merge clusters is based on the weighted average and average distance between all instances in the cluster.

To compute the distance between a pair of patterns, we use Squared Euclidean ($\sum (a_i - b_i)^2$), and to facilitate comparison between the output of HAC with the flat gold standard classes, we enforce flat clusters from the hierarchical clusters by applying a distance threshold t (a similarity cut-off point). This threshold determines whether an instance should be grouped within a cluster or not.

3.3.4 Topic Models

Topic modelling is an unsupervised approach, and as applied to a collection of documents as a means of discovering the prevalent themes that run throughout each document. These models learn both the salient topics (themes) throughout the text collection, and given these topics for the whole collection of documents, the model assigns topic probabilities to each document. For example, if topic models were applied over the day's newspaper, we would expect to discover topics based on all terms that appear in the paper. For example terms that may appear in the newspaper could be: *defender*, *serve*, *match*, *injury*, *goals*, *line-up*, *break*, *defeat*, *open*, *funding*, *education debt*, and *downturn*.

A distribution over these terms form 'topics', and a topic can be thought of as a collection of salient terms,[9] for example Topic 1 may include the salient terms: *serve*, *open*, *break*; Topic 2 may include: *match*, *injury*, *defeat*; Topic 3: *defender*, *goals*, *line*; and Topic 4 may have the terms: *funding*, *education*, *debt*, and *downturn*. These induced topics (a collection of semantically related terms), are also assigned with certain probilities over documents in the collection of documents. For example, a newspaper article discussing the French Open may have a probability assigment of Topic 1: 60%; Topic 2: 35%; Topic 3: 5%; and Topic 4: 0%. However an article

[9]All terms in a document collection are assigned probabilities, but only those with high probabilities are salient terms, and are the defining terms for the topic induced.

discussing the effects of the economic crisis may have the distribution of Topic 1: 4%; Topic 2: 0%; Topic 3: 0%; and Topic 4: 96%.

The models we employ are an implementation of *hierarchical Dirichlet processes* (HDP) as described by Teh *et al.* (2006), where the data determine the number of topics induced.[10] Our application of HDP defines a 'document' in a non-standard way: our 'document' is not a complete article, using the newspaper example, but a portion of the text surrounding our word of interest. This definition of a document in topic modelling has been applied in tasks such as word sense induction (WSI) (Lau *et al.* 2012), where a document is defined as the immediate context, or surrounding sentences of the target word. In WSI, the task is to learn to automatically discover the different senses of a given word. In the way that HDP is applied by Lau *et al.* (2012), the topic models represent the semantic context that define a sense, and the assignment of topics per word represent all the senses that can apply to that term.

We employ HDP in a similar way to Lau *et al.* (2012) for our work in identifying *-kan* alternations according to the kind of stem we have, as seen in Chapter 6. We aim to apply the topic models such that each topic represents the distribution of arguments allowable for that stem of that particular usage of *-kan*. The assignment of usages per stem (assignment of topics per 'document' or in our case, per stem) represents all the *-kan* alternations possible for that stem.

3.3.5 VerbNet

VerbNet is a hierarchical, domain-independent classification of English verbs based on Levin classes – it is a classification of subclasses of verbs that have members that share both syntactic and semantic similarities. We use VerbNet as a guide in constructing our own semantically coherent classes for Indonesian, in much the same way as Schulte im Walde (2006) creates her classes for German (see Section 2.6). This is unlike Sun *et al.*'s (2010) study who specifically create their gold standard data on syntactic alternations, and not only based on semantic similarity.

VerbNet has syntactic and semantic information, such as subcategorisation information, thematic grids, and selectional preferences. The version we use in this thesis is VerbNet 3.2 (Kipper *et al.* 2008), which comprises of over 5200 verbs senses, and over 3700 lemmas.[11]

3.3.6 Evaluation

The learning algorithms we employ in this thesis are all unsupervised methods. The evaluation of such methods in determining the quality of clusters found often

[10]The implementation we use employs Gibbs sampling and can be found at `http://www.cs.princeton.edu/~blei/topicmodeling.html`

[11]VerbNet is available from `http://verbs.colorado.edu/~mpalmer/projects/verbnet.html`.

use 'internal' criteria, such high intra-cluster similarity and low inter-cluster similarity (Sun 2012). However, for this study we use gold standard data against which we can compare the outcome of the systems we develop.

The evaluation metrics we employ in this comparison are *precision*, *recall*, and *F-score*, which is the harmonic mean of *precision* and *recall*. If C_S below represents a cluster (C) that is predicted by our system (S), and C_G is the grouping in the gold standard data (G) then, we can define Precision, Recall and F-score in the following way:

$$precision = \frac{(C_S \cap C_G)}{C_S}$$

$$recall = \frac{(C_S \cap C_G)}{C_G}$$

$$f = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

A variation on this measure that we also employ is called pairwise precision, recall, and F-score. Thus formulation is described in Menestrina *et al.* (2010) as follows:

$$PairPrecision(I,G) = \frac{|Pairs(I) \cap PairsG|}{|Pairs(I)|}$$

$$PairRecall(I,G) = \frac{|Pairs(I) \cap PairsG|}{|Pairs(G)|}$$

$$pF_1(I,G) = \frac{2 \times PairsPreciston(I,G) \times PairsRecall(I,G)}{PairsPreciston(I,G) + PairsRecall(I,G)}$$

In our implementation of this formula, we first convert all items in a cluster the gold clusters (G) and the induced clusters (S) into pairs. For example, if our experiment consisted of ingredients that included the items:

Herbs:	'rosemary', 'chives'
Legumes:	'peas', 'soybeans', 'lentils'
Leafy greens:	'spinach', 'bok choi', 'silverbeet'

The above Pairs(G) would consist of 7 items, namely:

# PAIRS	PAIRS
1	rosemary-chives
3	peas-soybeans; peas-lentils; soybeans-lentils
3	spinach-bok choi; spinach-silverbeet; bok choi-silverbeet

The pairs from the automated clustering (Pairs(I)[12]) are in turn converted to pairs and compared using the aforementioned formulation of $pP(I, G)$, $pR(I, G)$, and $pF_1(I, G)$ by Menestrina *et al.* (2010).

3.4 Discussion

The tools and methods presented here are not a substitute for critical thinking or linguistic analysis but rather a means to gather evidence. The ways in which the experiments are set up begin with a testable hypothesis, which requires linguistic knowledge, and knowledge about the limitations of the stochastic methods used, to design them.

There are indeed trade-offs with using these automated methods that sacrifice some depth of linguistic detail but depending on the linguistic question one asks, this will not diminish the quality of the study. The reason why we choose to employ stochastic methods for investigations such as determining word classes is because we want to be able to use as much data as we can, and not cherry-pick examples that may skew the way in which we may analyse our linguistic data.

[12] 'I' for induced clusters

Part II

Grammar Engineering

Chapter 4

Encoding Morphology

4.1 Introduction

In this chapter we report on various aspects of the implementation of the sublexical domain in Indonesian. In particular, we focus on voice marking, which is integral to the grammar in Indonesian. We implement this as an obligatory feature in the precision grammar, because it determines the argument linking and their surface realisation in a clause. Using Arka *et al.*'s (2009) implementation of the locative suffix *-i* as a stating point, we lay out our implementation of the suffix *-kan* given the properties described in Section 2.2.4. We implement the formal descriptions spelled out in terms of Lexical Functional Grammar in Arka (1993); Arka and Manning (1998); and Arka and Manning (2008), as well as changes to this formal description based on coordination evidence by Musgrave (2001). Both Arka (1993) and Musgrave (2001) give structural descriptions of the Indonesian sublexical word formation; Arka (1993) describes word formation in terms of the causatives, while Musgrave (2001) describes the attachment of non-subject clitics and their impact on alignment in Indonesian.

Finally, we investigate ways of imposing constraints on the application of the suffix *-kan* by devising stem types that will map out allowable morphosyntactic changes upon the affixing of *-kan* to certain stems.

4.2 Implementing Voice

Our implementation of voice is based on the findings of Arka (1993); Arka and Manning (1998); and Arka and Manning (2008). Unlike Tagalog, Indonesian does not have a spectrum of voice types as described by Foley (2008); it has an actor voice (AV) and an undergoer voice (UV) (in addition to a passive construction, see Section 2.2.3).

In this section we briefly outline the analysis upon which we base our implementation (again see Section 2.2.3 for more details on voice). We then present some

evidence by Musgrave (2001) that shows through nominal coordination that some of the underspecifications by Arka and Manning (2008) do not capture the data. We then update the initial analysis and finally present our implementation.

4.2.1 Arka and Manning's Solution (2008)

Before the binding evidence presented by Arka and Manning (2008), the verbal prefix *di-* was commonly analysed as a passive, in constructions such as Example (4.1b).

(4.1) [from (Vamarasi 1999:52)]

a. *Dia membuka pintu itu.*
3sg AV-open door that
"He opened the door."

b. *Pintu itu dibuka=nya*
door that di-open=3sgA
"The door was opened by him."

Vamarasi (1999) analyses *di-* as the derived passive counterpart to its corresponding *meN-* construction, as shown in Example (4.1a) and (b). Furthermore, all the constructions involving *di-*, for example the sentences in Example (4.2), were assumed to all have a unified analysis of passive.

(4.2) a. *Buku itu dibaca=nya.*
book this di-read=3sgA
"The book, he read."

b. *Buku itu dibaca (oleh) Ali.*
book this di-read. (by) Ali
"The book, read by Ali"

c. *Buku itu dibaca oleh=nya.*
book this di-read by=3sgA
"The book, read by him."

However, Arka and Manning (2008) argue that all constructions in Example (4.2) cannot be equivalent. Firstly, the third person enclitic agent *-nya* can bind with the reflexive pronoun *dirinya* in Example (4.3), suggesting that it is a core argument rather than an oblique. However, the agent *Ali* cannot bind the reflexive *dirinya* as shown in Example (4.4).

(4.3) *Dirinya tidak diperhatikan=nya*
3self NEG UV-care-KAN=3sgA
"(S)he looked after himself."
(Arka and Manning 2008:59)

(4.4) *?*Dirinya tidak diperhatikan Ali*
3REFL NEG PASS-care-KAN A
"Himself was not taken care of by Amir."
(Arka and Manning 2008:61)

Given this, and other evidence presented by Arka and Manning (2008), they conclude that the agent *Ali* in fact has a different status to the agent enclitic *-nya*, and that the construction in Example (4.3) can be equivalently analysed as undergoer or patient voice (Kroeger 1993; Foley 2008), where the agent in such a construction remains a core argument. On the other hand, Example (4.4) is an example of a passive construction, which disallows *Ali* and *dirinya* to be coindexed because *Ali* is an oblique, unlike *-nya*.

Furthermore, it is shown that the agent in a 'Pro-V construction' in Example (4.5) can bind with the reflexive subject, suggesting that the preverbal *kau* licenses an undergoer voice.

(4.5) [(Arka and Manning 2008:54)]

Dirimu mesti kau serahkan ke polisi.
2REFL must 2sg surrender to police
"You must surrender yourself to the police."

Arka and Manning (2008) present solutions by way of specifying the required lexical annotations to the verbal markers *meN-* and *di-*. The prefix *meN-* specifies that the *actor* is realised as the SUBJ.

When the verb is marked with *di-*, this signals that the more patient-like participant aligns with SUBJ, and that we have either a passive construction or an undergoer voice. Also the OBJ may be expressed as an enclitic to the verb or an independent noun phrase, as shown in Figure 4.1. In Examples (4.2b) and (4.2c), *Ali* or *-nya* can occupy the OBL position within a prepositional phrase (PP) headed by *oleh* "by". Also, *Ali* in Example (4.2b) can occupy the noun phrase slot (NP) in the phrase structure, with the functional annotation OBL.

In much the same way as *meN-* and *di-* dictate the linking of thematic roles to grammatical functions, it has been shown that in some languages pronominal clitics have this same function, for example Tukang Besi (see Donohue (2004)), where the presence or absence of pronominal clitics determines whether the clause has an ergative or accusative linking. In Indonesian, the AV(*meN-*) signals an accusative linking, shown in Example (4.6).

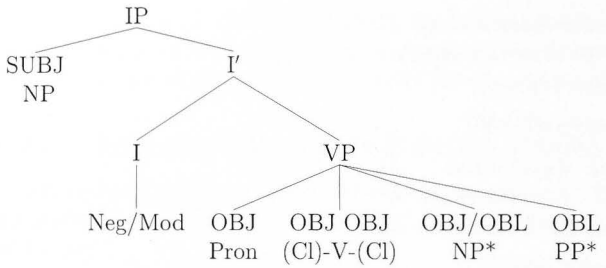


Figure 4.1: Phrase Structure per Arka and Manning (2008)

(4.6) *Saya sudah membaca buku itu.*
1sg already AV-read book that
"I have already read the book."

(4.7) *Saya baca buku itu.*
1sgread book the
"I read the book."

Chung (1978) also shows that there are also bare AV verbs, which she says is possible to have been arrived at through *meN-* deletion, but the word order remains unchanged, as shown in Example (4.7). Example (4.6) exhibits an S V O word order, as does Example (4.7).

There are two ways to signal an 'ergative' linking: (1) morphologically with *di-V-nya* (Arka and Manning 2008); and (2) through word order – Patient Agent Verb. This latter construction Musgrave (2001) calls the Pro-V construction, and is named object shifting by Chung (1978), where the agent or actor is attached directly to the left of the verb. Chung (1978:343) describes this agent as "giving the appearance of having cliticised to the left of the verb", as shown in Example (4.8), where *sudah* "already" is not able to inserted between the pronoun *kami* "we", and the verb *read* "baca".

(4.8) [from Chung (1978:343)]

a. *Buku itu sudah kami baca.*
book that already 2pl.EXCL read
"The book, we already read."

b. **Buku itu kami sudah baca.*
book that 2pl already read

FOR: "The book, we already read."

Musgrave (2001:76) also notes that with the Pro-V construction the patient can be right dislocated, for example Example (4.9). Therefore without temporals, modals or negation, this construction is indistinguishable from a bare AV clause shown in Example (4.9).

(4.9) [from Chung (1978:342)]

Bisa kami terbangkan layangan itu.
Can 2pl TER-fly-KAN kite this
"We can fly the kite."

However, in the construction in Example (4.10), the agent is in fact cliticed to the verb, and we analyse this kind of construction in the same way as Example (4.8a) for implementation, even though *ku-* "1sg" is a pronominal clitic, while *saya* "1sg" is a full pronoun. We discuss this preverbal position in more detail in Section 4.2.2.

(4.10) *Buku itu ku=baca.*
book this 1sg=read
"This book, I read."

The analysis of *di-V-nya* as being distinct from the passive *di-* account for the binding evidence of reflexives. Also, the phrase structure in Figure 4.1 account for the other undergoer voice constructions that are signalled with pronominal proclitics, such as Example (4.10). However, Musgrave (2001) presents evidence from nominal coordination that compels us to update the structure presented in Figure 4.1, which we discuss in the following section.

4.2.2 Coordination Evidence from Musgrave (2001)

In this section we present coordination evidence from Musgrave (2001) that shows that post-verbal object clitics should not be encoded morphologically, as suggested in Figure 4.1. Also, given the restriction of the preverbal agents in Indonesian, that the syntactic OBJ position in Figure 4.1 would only lead to generating constructions that are not allowable in the language.

(4.11) [from Musgrave (2001:92)]

a. *Saya mencintaimu dan ibu=mu*
1SG AV-love-I=2sg and mother=2sg
"I love you and your mother."

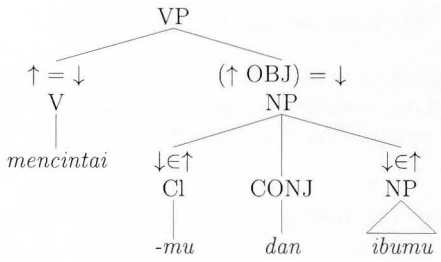


Figure 4.2: Phrase structure for VP: *love you and your mum*

b. *Saya mencintai =nya/ saudara=mu dan ibu=mu*
1SG AV-love-I =3sg/ sibling=2sgand mother=2sg
"I love her/your sister and your mother."

c. *Saya mencintai kamu dan ibu=mu*
1SG AV-love-I 2sg and mother=2sg
"I love you and your mother."

We see from Musgrave's (2001) example in Example (4.11a), a full noun phrase *ibumu* "your mum" can coordinate with a patient clitic *-mu* "you", suggesting that they should be treated in the same way as independent pronouns and common nouns. Therefore, we implement the object enclitics *-mu* and *-nya* in Examples (4.11a) and (4.11b) structurally in the same way as Example (4.11c). Given that the enclitic *-mu* "2sg" can coordinate with the full NP, we allow the clitic to share a node with the NP *ibumu* "your mum", as shown in Figure 4.2.

While the patient clitic *-nya* in Example (4.11b) can coordinate with a NP, the enclitic *-nya* in Example (4.12) cannot when the verb is prefixed with *di-*.

(4.12) [from (Musgrave 2001:92)]

**Siti dilihatnya dan ibunya.*
Siti UV-see=3sgA and mother=3sg
(For: "Siti was seen by her and her mother.")

In Example (4.12) *-nya* cannot coordinate with other noun phrases, however there is no restriction on other kinds of nominals being in a coordination construction when the main verb is prefixed with *di-* as shown in Sentences 4.13 and 4.14

(4.13) [from (Musgrave 2001:92)]

Buku itu dibaca mereka dan kita.
book that PASS-read 3pland 1pl.INCL
"The book was read by them and by us."

(4.14) *Buku itu dibaca guru dan mahasiswa.*
book that PASS-read teacher and student
"The book was read by the teacher and by the student."

(Musgrave 2001:92)

Also, preverbal coordination even between pronouns and pronominal clitics are disallowed, as shown in Example (4.15).

(4.15) [from Musgrave (2001:90)]

**Anjing itu saya dan dia pukul*
dog that 1sgand 3sghit
FOR "I and he hit the dog."

This is why unlike Arka and Manning (2008), we have not implemented a syntactic position before the verb.

Musgrave (2001) argues for a sublexical agent position after the verb. The evidence Musgrave (2001) offers for such a structure, is that agents in this object position cannot be separated by adjuncts, such as *dulu* "before" in Example (4.16). He shows that for AV-marked verbs this is possible, as shown in Example (4.17). However, Arka (1993:68) states that there is no adjunct insertion point directly after the verb, regardless, and contrary to Example (4.17), we see from the object *mobil* "car" also cannot be separated from the verb *membeli* "buy" in Example (4.18).

(4.16) from Musgrave (2001:100)

*Film itu dilihat dulu *(oleh) Umar*
film that PASS-see before by U
"That film was seen previously by Umar."

(4.17) from Musgrave (2001:100)

Saya membeli sekarang buku itu
1sg AV-buy now book that
FOR: "I bought the book just now."

(4.18) from Arka (1993:68)

**Ia membeli kemarin mobil*
3sg AV-buy yesterday car
FOR: "He bought yesterday a car."

Although, we could not find examples in Wikipedia akin to Example (4.17), which allows a temporal adverb in this post-verbal position, we can only conclude that adjunct placement is not a satisfactory test for determining how we should implement the postverbal agent of the *di*-marked verb.[1] However, if we examine Examplea (4.16), it is possible to have coordinated NP agents in this passive construction, as we had seen in Examples (4.13) and (4.14), suggesting that this agent should not be encoded sublexically.

4.2.3 An Updated Solution for Implementation

In the preverbal position there are no syntactic slots in our implementation; no common nouns are allowed preverbally, as shown in Figure 4.3, only pronominals, and pronoun substitutes, which are encoded as sublexical rules.

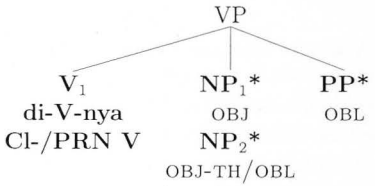


Figure 4.3: Phrase structure for remodeled VP

The **Cl-V** and **PRN V** are the constructions Musgrave (2001) call Pro-V constructions. Under the label **PRN V** in Figure 4.3, we also encode pronoun substitutes, and $\mathbf{V}_1$ encodes all other verb forms that are not UV, including bare actives and AV forms and the passive verb. Also, under the label $\mathbf{NP}_1$ is the syntactic position for enclitics. Figure 4.4 shows details of the sublexical rules for the implementation of Figure 4.3.

The preverbal agents are encoded sublexically as shown Figure 4.4. The label VOICE encodes the AV and passive. The AV is ordinarily marked with the *meN-* prefix, but as suggested by Chung (1976), these AV verbs can undergo *meN-* deletion.

[1] Although we could not find examples to support Example (4.17), this kind of evidence is difficult to search for using regular expression searches, without tagged corpora such as a treebank.

$$
\begin{array}{lll}
\text{V} & \rightarrow & \{\ \text{PRE-V}\ \ \text{V'}_{stem}\ \mid\ \text{di-V-nya}\ \} \\
\text{PRE-V} & \rightarrow & \{\ \text{VOICE}\ \mid\ \text{PreVn}\ \} \\
 & & \downarrow = (\uparrow \text{OBJ}) \\
 & & (\uparrow \text{VOICE-TYPE}) = \text{UV} \\
\text{PreVn} & \rightarrow & \{\ \text{Cl}\ \mid\ \text{Pron}\ \mid\ \text{PreName}\ \} \\
\text{di-V-nya} & \rightarrow & \text{DI}\ \ \text{V'}_{stem}\ \ \text{Cl} \\
 & & \downarrow = (\uparrow \text{OBJ})
\end{array}
$$

The node PreName represents a proper noun (personal name), and although PreName may also include honorifics, it is not a full noun phrase, but it is a pronoun substitute, and it is a politeness strategy to avoid the second person pronoun (see Section 2.2.2 for more on pronominals and pronoun substitutes).

Figure 4.4: Sublexical rewrite rules

(4.19) *Mobil saya Pak Ali beli.*
car 1.sgHON A buy
"My car, you will buy." (Pro-V, addressing Ali)
"My car, Ali will buy." (object fronted, bare active)

Proper names are allowed preverbally, but they are not full NPs and their usage is very restricted as pronominal substitutes in a Pro-V construction, as shown in Example (4.19) (see sec:pronouns-ind for more on pronominal substitutes).

Example (4.19) can be interpreted in two ways: as a Pro-V construction, or an object focused construction where the object is shifted in a bare AV sentence.

In the preverbal position shown in the sublexical rules in Figure 4.4, we see that PRE-V expands to a choice between VOICE (the choice between *meN-* or *di-*) or a PreVn, which is a preverbal agent. This correctly predicts forms such as Example (4.20) should be ungrammatical, but it also reflects the fact that this Pro-V construction determines argument linking, and therefore predicts that a sentence such as Example (4.20c) should be ungrammatical. This sentence has a verb with no voice marker, and it does not conform to the word order required for a Pro-V construction, and therefore analysing this as a *meN-* deleted AV would be pragmatically strange and syntactically incomplete.

(4.20) Predicts ungrammaticality of:

a. **Mobil itu kumembeli.*
car that 1sg=AV-buy
FOR: "This car, I bought."

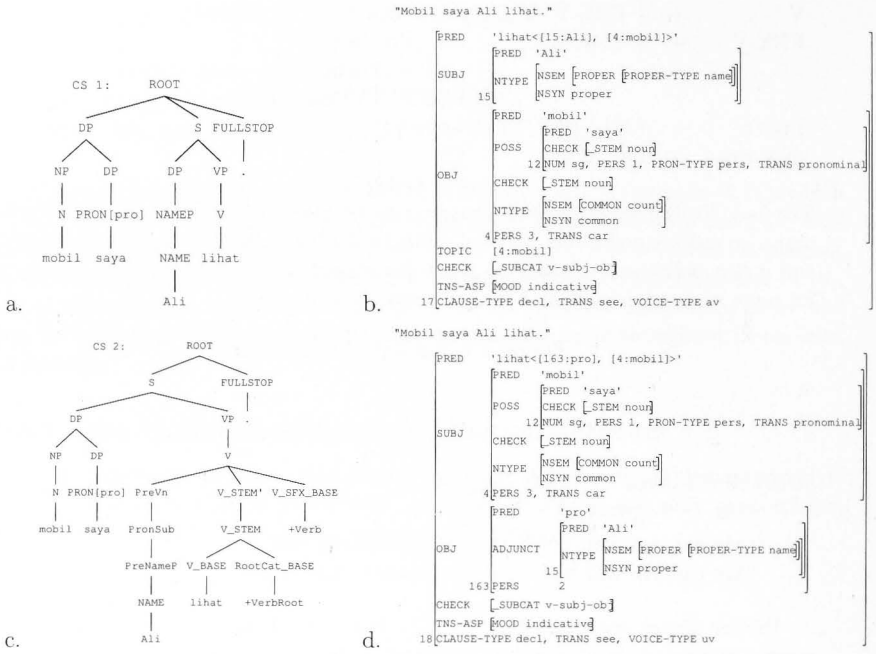


Figure 4.5: Parses for Example (4.19)

b. **Kumembeli mobil itu.*
1sg=AV-buy car that
FOR: "I bought this car." (postposed subject)

c. **Mobil itu beli.*
car that buy
FOR: "The car was bought."

In order to ensure that VOICE is obligatorily encoded in the f-structure, our common definition of the VERB defined in the **template** in Figure 4.6, which applies to all verbs in the lexicon has an existential constraint on line 2.[2] This constraint states

[2]Templates are a way that collection of functional descriptions or LFG equations can be given an alias, so that this alias can represent this set of functional descriptions. See Dalrymple *et al.* (2004).

```
1   VERB(_G) = { "common to all verbs"
2                (^ VOICE-TYPE)
3                (^ CHECK _CONSTR) ~= derivedn
4                (^ TRANS) = _G
5              |
6                (^ CHECK _CONSTR) =c derivedn
7                (^ TRANS) = _G
8              }.
```

Figure 4.6: Verb template

that the f-structure associated with its mother node must have the feature VOICE-TYPE, unless it is a derived nominal.[3] This constraint, however, does not assign a value, and it must be determined elsewhere, either from a voice marker or constructionally. This is how we encode the symmetrical *Philippine-type* nature of voice in Indonesian (Arka and Manning 2008; Foley 2008).

The linking of arguments can be signalled by the AV(*meN*) prefix on the verb, or via word order with the bare verb: S V O, both resulting in an accusative linking and O S-V (or a Pro-V construction) triggers an ergative linking and encodes a UV VOICE feature, as does the *di-V-nya* construction.

In addition to being able to model why coordination is permitted post-verbally and not preverbally, the phrase structure we implement also predicts that Example (4.21) should be ungrammatical whereas the phrase structure proposed by Arka and Manning (2008), with the encoding of proclitic *ku-* as triggering undergoer voice, as we have here, would allow this.

(4.21) Predicts ungrammaticality of:

**Buku itu Ali kubelikan.*
book this A 1.sg-buy-KAN
FOR: "This book, I bought Ali."

The coordination facts presented in Section 4.2.2 show why we analyse the enclitics as sharing a node with NP objects for implementation, as illustrated in detail in

[3]Sneddon *et al.* (2010) suggest that the nominalisers *peng-* and *pe-* derived nouns that encode a *doer of an action* and the *the result of an action* depicted by the verb, respectively. Although Sneddon *et al.* (2010) do not analyse them as such, it suggests some encoding of the voice properties even in nominalised forms, but this has been left for future work because we had not yet investigated this fully.

	Stem	Stem+*i*
Type 1	SUBJ	SUBJ, OBJ
Type 2	SUBJ, OBJ, OBJ2	SUBJ, OBJ, OBL
Type 3	SUBJ, OBJ	SUBJ, OBJ
Type 4	SUBJ, OBJ	SUBJ, OBJ { OBJ2 | OBL }

Figure 4.7: Summary of the types of changes imposed on the argument structure of the *-i* verb by Arka *et al* (2009).

Figure 4.3. Also restructuring the preverbal agent as being part of the sublexical domain prevents ungrammatical sentences such as Example (4.21).

4.3 The suffix -kan

As we had seen in Section 2.2.4, the variation of changes to the predicate-argument structure of *kan*-affixed verbs are numerous. This section focuses on the implementation of the *-kan* suffix, and the sublexical rules required to account for the changes in argument structure to *kan*-affixed verbs, as discussed in Section 2.2.4. Our implementation of *-kan* follows the implementation of *-i* (Arka *et al.* 2009), which is discussed in the following section.

4.3.1 The implementation of -i as a model for -kan

The suffix *-i* is described as a locative applicative (Arka *et al.* 2009), and like *-kan* does not always change the number of arguments an affixed verb takes, but can serve as an aspectual marker. Arka *et al.* (2009) characterises *-i* as a progressive marker, and Son and Cole (2008) regards *-kan* as applying a resultative reading, when its function is not to applicativise.

Arka *et al.* (2009) identify 4 kinds of applicative constructions imposed by *-i*, summarised in Figure 4.7. These changes to the argument structure are described in terms of the transitivity profile of the verb that *-i* attaches to. Examples of each type are shown in Examples (4.22) to (4.25). Type 1 constructions, such as Example (4.22), are defined in the lexicon as intransitives, and the suffix *-i* forms transitive verbs. Type 2 constructions, such as Example (4.23), allow the *theme* (*minum* "drink") to be demoted to OBL status. For certain transitives the affixing of *-i* makes no changes to the subcategorisation frame of the verb, for example Type 3 constructions in Example (4.24). Type 4 optionally demotes a direct OBJ to OBL, as seen in Example (4.25).

(4.22) [Type 1 from Arka *et al.* (2009)]

a. *Mangga yang besar jatuh ke rumahnya.*
mango that big fall to house=3.sg
"A big mango fell onto his house."

b. *Mangga yang besar menjatuhi rumahnya.*
mango that big AV-fall-I house=3.sg
"A big mango fell onto his house."

(4.23) [Type 2 from Arka *et al.* (2009)]

a. **Engkau menyuguh minum lezat kepada aku.*
2.sgAV-serve drink tasty to 1.sg
FOR: "You served a very tasty drink to me."
(Both complements must be direct arguments.)

b. *Engkau menyuguhi aku minum lezat*
2.sgAV-serve 1.sgdrink tasty
"You served me a very tasty drink."

c. *Engkau menyuguhi aku dengan minum lezat*
2.sgAV-serve 1.sgwith drink tasty
"You served me a very tasty drink."

(4.24) [Type 3 from Arka *et al.* (2009)]

a. *Ia memukul saya*
3.sgAV-hit 1.sg
"S/he hit me."

b. *Ia memukuli saya*
3.sgAV-hit-I 1.sg
"S/he hit me."

(4.25) [Type 4 from Arka *et al.* (2009)]

a. *Air itu sedang mengalir ke sawah.*
water that in.progress AV-flow to rice.field
"The water is flowing to the rice field."

b. *Dia mengaliri sawahnya dengan air itu.*
3.sgAV-flow-I rice.field with water that
"S/he flooded his/her rice field with water."

c. *Dia mengaliri sawahnya air itu.*
3.sgAV-flow-I water that
"S/he flooded his/her rice field with water."

```
1 APPL_I =
2       {(↑ PRED) = 'V_Appl_i <(↑ SUBJ) (↑ OBJ) %PRED3>'
3           ↑\PRED\GF = ↓\PRED\GF                          Type 1: IntrRoot → Vtr
4                       { (↓ SUBJ) = (↑ SUBJ)
5                       (↓ OBL-LOC)=(↑ OBJ)        |
6                       (↓ SUBJ) = (↑ SUBJ)
7                       (↓ OBL-LOC) = (↑ OBJ)      Type 2: TrRoot → Vtr
8                       (↓ OBJ)= (↑ OBL-INST)
9                       (↑ OBL-INST CASE)=c obl-inst  |
10                      (↓ SUBJ) = (↑ SUBJ)
11                      (↓ OBJ)=(↑ OBJ)            Type 3: TrRoot → Vtr
12                      (↑ TNS-ASP PROG)=+
13                      ~(↑ OBL-INST) "just for the iterative meaning of -i" }
14                      (↓ PRED)=(↑ PRED ARG3)        |
15          (↑ PRED) = 'V_Appl_i <(↑ SUBJ)(↑ OBJ) (↑ OBJ2) %PRED4>'
16          ↑ \PRED\GF= ↓ \PRED\GF
17                      (↓ SUBJ) = (↑ OBJ)
18                      (↓ OBL-LOC) = (↑ OBJ)      Type 4: IntrRoot → Vtr
19                      (↓ OBJ)= (↑ OBJ2)
20                      (↓ PRED)=(↑ PRED ARG4) }
21      (↑ APPLICATIVE)= +.
```

Figure 4.8: Template from Arka *et al* (2009) for applicative *-i* construction

Given this description of the suffix *-i*, Arka *et al.* (2009) attach the information in Figure 4.8[4] to the *-i* construction. What the APPL_I template aims to capture is that there are two resulting subcategorisation frames associated with the applicative *-i*. These two resulting subcategorisation frames are encoded in lines 2 and 15. The first 3 types define a resulting predicate that takes two arguments, and the last type results in a three place predicate.

For the *-kan* implementation, like *-i*, we take a predicate composition approach as described in Section 2.3.1 (under *Argument Structure*) we assume that *-kan* is an incomplete predicate with annotated instructions on how the argument structure is affected.

4.3.2 The implementation of -kan

From the variations of the alternation to the a-structure imposed by *-kan* discussed in Section 2.2.4, we summarise the possible changes to the subcategorisation information of the *kan*-affixed verb in Table 4.1. These syntactic changes are grouped into 5 types, according to the kind of arguments the verb takes, as it is defined in the lexicon, as well as the resulting *kan*-affixed verb.

[4]This template employs an operator known as the 'restriction operator' (signalled by '\'), which restricts the information that is propagated in the syntactic structure. See Section 4.3.2 for more details.

Example #	Example	Stem	Stem+KAN
Type 1. Benefactive applicative:			
2.37 2.36b	*jahit* "sew"	SUBJ, OBJ_i	SUBJ, OBJ1, $OBJ2_i$
Type 2a. Causative:			
2.37 2.36a	*jahit* "sew"		
2.38	*balut* "wrap"	$SUBJ_i$, OBJ	SUBJ, OBJ, OBL_i
2.41 2.42	*muat* "load"		
Type 2b. Causative:			
2.40	*tikam* "stab"	SUBJ, OBJ_i	SUBJ, OBJ, OBL_i
Type 3. Optional -kan:			
2.39	*ikat* "tie"	SUBJ, OBJ	SUBJ, OBJ
Type 4. Benefactive applicative:			
2.43 2.44	*beri* "give"	SUBJ, OBJ1, $OBJ2_i$	SUBJ, OBJ $_i$, OBL
Type 5. Causative:			
2.45	*datang* "arrive"	SUBJ	SUBJ, OBJ

Table 4.1: Variations to Subcategorisation Information for *kan*-affixed verbs

Based on these five descriptions, we define the template for the incomplete predicate *-kan* in Figure 4.9. Each of the top-level disjunctions in the definition (|, 'or') represent each type in Table 4.1. As mentioned by Kroeger (2007), many of the peripheral interpretations of *-kan* such as the instrumental alternation (see Section 2.2.4), are in fact a kind of causative construction, which we reflect in the grouping of types in Table 4.1.

The restriction operator '\' on lines 4, 11, 20, 25, and 32 is defined formally by Kaplan and Wedekind (1993:198), its application results in the restricting of information being propagated in a given f-structure.[5] In terms of XLE, the restriction implementation allows "f-structures and predicates to be manipulated and controlled in a detailed fashion" (Butt *et al.* 2003:96).

The first *-kan* choice defined in lines 2 to 6 in KAN-PRED encodes Type1 from

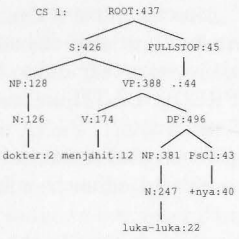
[5]An example of how the restriction operator affect the f-structure is given by Kaplan and Wedekind (1993). This example shows the effect '\' in $f\backslash$SUBJ has on an f is as follows:

$$f = \begin{bmatrix} \text{PRED} & \text{kick} \\ \text{SUBJ} & \text{John} \\ \text{OBJ} & \text{ball} \end{bmatrix}$$

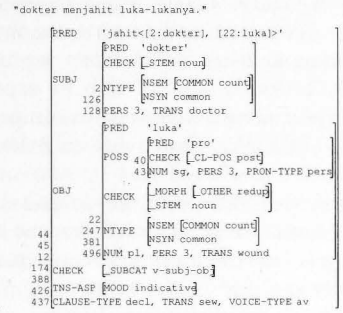
$$f\backslash\text{SUBJ} = \begin{bmatrix} \text{PRED} & \text{kick} \\ \text{OBJ} & \text{ball} \end{bmatrix}$$

```
1   KAN-PRED( _P) =
2  {   "TYPE 1 eg jahit (sew)"
3      (^ CHECK _SUBCAT) =c v-subj-obj
4      ^\PRED\OBJ-TH = !\PRED
5      (^ PRED) = '_P<%ARG1 (^ OBJ-TH)>'
6      (! PRED)=(^ PRED ARG1)
7   |  "TYPE 2 eg balut (wrap), tikam (stab), jahit (sew)..."
9      (^ CHECK _SUBCAT) =c v-subj-obj
10     (! PRED)=(^ PRED ARG1)
11     ^\PRED\OBL = !\PRED
12     (^ PRED) = '_P<%ARG1 (^ OBL)>'
14     {
15       (^ OBL CHECK _PFORM) = 'ke'
16     | (^ OBL CHECK _PFORM) = 'dengan'
17     }
18  |  "TYPE 3 eg ikat (tie/chord)"
19     (^ CHECK _SUBCAT) =c v-subj-obj
20     ^\PRED = !\PRED
21     (^ PRED) = '_P<%ARG1>'
22     (! PRED)=(^ PRED ARG1)
23  |  "TYPE 4 eg beri (give)"
24     (^ CHECK _SUBCAT) =c v-subj-obj-th
25     ^\PRED\OBL = !\PRED\OBJ-TH
26     (^ PRED) = '_P<%ARG1>'
27     (! PRED)=(^ PRED ARG1)
28     (^ OBL) = (! OBJ-TH)
29     (^ OBL CHECK _PFORM) = 'kepada'
30  |  "TYPE 5 eg datang (come)"
31     (^ CHECK _SUBCAT) =c v-subj
32     ^\PRED\OBJ = !\PRED
33     (^ PRED) = '_P<%ARG1 (^ OBJ)>'
34     (! PRED)=(^ PRED ARG1)
35  }.
```

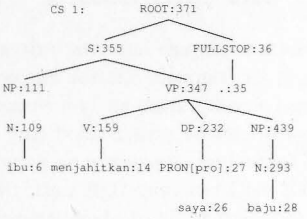
Figure 4.9: Template for incomplete predicate *-kan*



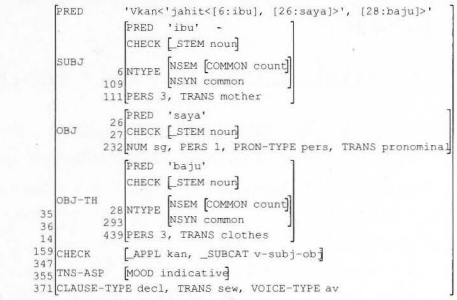
a.



b.



c.



d.

a&b. *dokter menjahit luka-lukanya*
doctor AV-buy wound-wound=3.sg
"the doctor sewed his/her wound"

c&d. *Ibu menjahitkan saya baju.*
mother AV=sew-KAN 1.sg shirt
"Mother sewed me a shirt."

Figure 4.10: c-stucture and f-structure for Type 1

Table 4.1. The resulting *kan*-affixed PRED (on line 5) copies the original PRED as defined in the lexicon with its argument list, represented as %ARG, and appends another argument 'OBJ-TH' in the newly composed predicate. This results in an argument list of length three: ⟨ SUBJ, OBJ, OBJ-TH ⟩. This can be seen in the f-structure in Figure (4.10d), where the complete f-structure subcategorises for three arguments SUBJ, OBJ, and OBJ-TH, for the verb *menjahitkan* "sew X Y", while the corresponding non-*kan* verb has two arguments, as seen in Figure (4.10b).

Line 3 checks the kind of verb we expect with (^ CHECK _SUBCAT) =c v-subj-obj; we expect a transitive verb to compose with KAN-PRED at this disjunct. The !\PRED (↓ \PRED) indicates that only the PRED value is restricted, and so all other f-structures are unaffected, and ^\PRED\OBJ-TH (↑ \PRED\OBJ-TH) indicates that these values are redefined in the composed resulting f-structure.

Lines 15 and 16 for Type 2 define the kind of preposition that needs to head the OBL. This is because all prepositional phrases are considered adjuncts unless they are lexically specified by the verb.

For Type 3, which renders *-kan* optional, we analyse *ke anjing* "to the dog" as an adjunct because it is optional, as seen in Example (2.39a), repeated in Example (4.26).[6]

(4.26) *Optional PP:*

a. *Dia mengikat tali itu.*
3.sg AV-tie rope that
"S/he tied the rope."

b. *Dia mengikat tali itu ke anjing*
3sg AV-tie-KAN rope that to dog
"S/he ties the rope to the dog."

For the definition of this *optional -kan*, line 21 shows that there is no changes or additions to the argument list, unlike Type 1, which appends an OBJ-TH, as shown in line 5. The only restricted items are the PRED values for the verb, we can redefined a *kan*-composed predicate, and apart from that renders no other changes.

[6]We maintain the *optional* label applied by Son and Cole (2008), for the *ikat* "tie" in Example (2.39), as shown in the example below.

Dia mengikat(-kan) tali itu ke anjing
3sg AV-tie-KAN rope that to dog
"S/he ties the rope to the dog."

But the behaviour of the *mengikat* "tie" and *mengikatkan* "tie to" are not completely identical, and with the *kan*-affixed verb, the PP locative, *ke anjing* "to the dog", has to at least be assumed if not expressed, although this is not true for *mengikat*, the non-*kan*-affixed verb.

Each of the 5 types summarised in Table 4.1 are parsed alongside their non-*kan* counterparts and shown in Appendix B.

Although this implementation captures the data presented in Section 2.2.4, there are some major issues, mainly regarding overgeneration. This implementation allows all *kan*-affixed verbs to participate in all the alternations, as long as one of their lexical definitions matches the _SUBCAT value required for that definition. For example this implementation allows the verbs *jahit* "sew", and *muat* "load" to apply optionally, although it is not attested.

Another problem is that not all stems that take the *-kan* suffix are verbs. For example, *ikat* "tie" is primarily a noun,[7] which obtains its subcategorisation information if affixed with AV.

In Arka's (1993) definition of the causative *-kan*, he maps out morphosyntactic and semantic changes, within and across word classes for a small number of stems. In the next section, we do a more extensive survey to help us define types for *-kan* to help mitigate overgeneration and overapplication of *-kan* in the lexicon. Our goal is to define classes of stems and that map out the possible *-kan* alternations that only apply to them.

4.4 Extending the Lexical Coverage for -kan

In the previous chapter we see how a small number of stems are implemented to account for the alternations imposed by *-kan*. However, not all stems behave in the same way as each other when affixed with *-kan*. Although, the behaviour of *-kan* has been shown to be rather varied, at times adding an argument, at times removing an argument, and at other times not affecting the number of arguments at all (Kroeger 2007; Son and Cole 2008) (see Section 2.2.4), not all of the stems that we show participate in all of these alternations. For example, *ikat* "tie" can participate the *Optional -kan* construction as shown in Table 4.1, in Section 4.3. However, this alternation does not apply to the stem *tikam* "stab" – it does not optionally allow *-kan* to be attached. In this section, we investigate ways that we can manually cluster the variations seen with the combining of *-kan* and a range of stems, so that we can define subclasses of stems that all exhibit the same alternations. The way we cluster these like verbs is by tracking their semantic and syntactic changes relative to *kan* by decomposing and identifying specific semantic and syntactic changes that are undergone by the stem when *-kan* is affixed. This method of lexical decomposition, in order to find these verbs than behave in the same way under the same morphological conditions, is explained in Section 4.4.2.

We choose 100 stems in total that are labelled as either noun, adjective, or verb in the *Kamus Besar Bahasa Indonesia* (KBBI – 'The Big Indonesian Language Dic-

[7]The primary sense of *[tie/cord]ikat* is a noun according the KBBI *Kamus Besar Bahasa Indonesia* "The Big Indonesian Dictionary", the official Indonesian language dictionary (Sugiono 2008)

tionary') (Sugiono 2008).[8] In our discovery of clusters and their variations according to stems, we take a data-driven approach, which is agnostic about the nature of *-kan*. Specifically, we make no specific assumptions about whether there are two homophonous *-kan* affixes as suggested by Kroeger (2007) or a single *-kan* as suggested by Son and Cole (2008), but allow this to fall out from the data. This chapter aims to extend the variation shown by Arka (1993) beyond the causative *-kan* (see Section 2.2.4). However, we aim, not to just concentrate on one kind of phenomenon as a result of *-kan*, but the spectrum of variation imposed by the combination of the stem and *-kan*.

Our goal in this corpus study is to be able to group stems that share the same alternation when affixed with *-kan*. We aim to define these groups of *-kan* alternations as *types* that we employ in the lexicon in order to restrict the kind of alternations a particular stem of a particular *type* can participate in.

4.4.1 Assumptions

As pointed out by Jackendoff (2002), for English at least, the correspondence between the derived verbs from nouns are more often arbitrary. For this reason meaning changes are encoded in the semantic decomposition for noun stems.

We also assume that the sense of the word that is listed first in the Indonesian dictionary KBBI (Sugiono 2008) is its primary sense and therefore we take this definition for the word's meaning and word class. While we did not always agree with the relative ranking of the senses in the dictionary (e.g. listing 2 is more common in colloquial Indonesian than listing 1, in the example in Figure 4.11), we remain faithful to the KBBI listing.

pusing **1** *v* to go to and fro **2** *a* ill (usually with a headache) **3** *a* feeling unbalanced as though the surroundings are whirling around **4** *a* bewildered

Figure 4.11: Dictionary entry for *pusing* in the KBBI, simplified and translated from Indonesian

4.4.2 Embarking on Manual Text Analysis

This section outlines the method used in collecting information upon which we derive our analysis and characterisation of how *kan* affects the argument structure of verbs. The method involves collecting evidence of usage from a text collection, and the methodology can be summed up as a *corpus-driven approach*.

[8]The KBBI is the official dictionary of the Indonesian language released by the Indonesian government.

> A corpus-driven approach involves a bottom-up methodology, beginning by selected unedited examples form the corpus, identifying their shared and individual features, and only then grouping them for the purpose of lexicographic representation
>
> (Krishnamurthy 2008:231)

Although this corpus-driven approach may seem straightforward and self-explanatory, once one embarks on the process, there are many details to fill in. For example discovering what *features* are important and noteworthy for our purpose; what process is undertaken when selecting *unedited examples from the corpus*; what are the factors in determining what is meant by *shared and individual*. In the following sections, we detail the steps we take in this process.

Nouns		**Adjectives**	**Verbs**
analogi (analogy)	administrasi (administration)	abadi (eternal)	acuh (heed)
ajar (instruction)	aplikasi (application)	agung (sublime)	baca (read)
asumsi (assumption)	belanja (expenses)	asing (strange)	bangun (get up)
buku (book)	darat (land)	biasa (common)	bawa (bring)
didih (boiling)	ekspresi (expression)	berani (audacious)	beri (give)
gambar (picture)	gelembung (bubble)	cemar (dirty)	buat (make/do)
hipotesis (hypothesis)	ikat (tie)	cengang (amazed)	hadir (be present)
injeksi (injection)	instalasi (installation)	cerdas	dengar (hear)
janji (promise)	letak (position)	goyah (unstable)	hidup (live/be alive)
kerja (work/labour)	kait (hook)	haram (prohibited)	jatuh (fall)
legalisasi (legalisation)	kumandang (echo)	kecewa (disappointed)	kenang (think of)
lokasi (location)	cerita (news)	lanjut (protracted)	lulus (go through)
mimpi (dream)	maklumat (declaration)	leceh (worthless)	mandi (bathe)
nasionalisasi (nationalisation)	mula (beginning)	lunak (soft)	masuk (enter)
paten (patent)	pikir (idea)	murni (pure)	mati (die)
penjara (jail)	publikasi (publication)	mutakhir (up-to-date)	minggir (to put aside)
pukul (hit/blow)	percik (stain)	padu (compact)	pecah (break)
radiasi (radiation)	pusat (centre)	populer (popular)	paksa (force)
sesal (regret)	rumah (house)	pilu (sad/moved)	pusing (be concerned)
susu (milk)	sisa (remainder)	remeh (unimportant)	serah (surrender)
sewa (lease)	sarang (web/net)	salah (wrong)	singkir (get out of way)
tempat (place)	tanya (question)	subur (fruitful)	susup (duck down)
tumpu (foothold)	titah (a blow)	takjub (surprised)	terjemah (translate)
wakil (proxy)	umpama (example)	teguh (strong)	tewas (perish)
		terang (clear)	timpa (hit)
		unggul (excellent)	
		jengkel (annoyed)	

Table 4.2: 100 stems with first sense determining the categorisation of word class

In this process, we aim to find a way in which we can explain the syntactic and the semantic differences between two related morphological contexts in term of the

stem, namely when the stem is prefixed with only with MEN, as shown in 1, and when this form is additionally suffixed with *-kan*, as shown in 2.

1. MEN+stem
2. MEN+stem+KAN

We had identified 735 stems with attested *-kan* usage, and randomly chose a subset of 100 stems to investigate, shown in Table 4.2. For our task, we are not interested in the occurance of the specific surrounding lexemes of our target stem, but instead aim to infer patterns of subcategorisation from the text, for the instances of the stems we find in our text. We use the Indonesian Wikipedia as our text source (see Section 3.3.1).

We coarsely map out the variation of arguments for each target stem in the following way. We search for sentences that contain the stem in pattern 1 (meN+stem) in our text collection. We restrict the number of sentences we analyse to 25-50. For each of the sentences we construct a verb template or verb frame that describes the subcategorisation frame for that verb. For example, take the adjective stem *lunak* "soft". This adjective when prefixed with *meN* gives us *melunak* "soften", and searching for sentences with the predicate *melunak* "become soft" returns instransitive sentences for all sentences we find, and so our verb frame would be: NP_a, for this particular A(djective) stem. We then compose our semantics around these immutable elements – NP and A. We then repeat this process for pattern (2) for the same stem.

The main focus of this investigation is not on semantics. However, in our effort to be systematic in our analysis, we rely on semantic tools formulated in a couple of models of lexical semantics, namely Natural Semantic Metalanguage and Conceptual Structure. Our aim is to be able to decompose each verb into smaller meaning components based on the stem, in order to group similar stems together.

4.4.3 A Smorgasboard of Semantics

For this task, we require a semantic formalism that will work as scaffolding around the stem, so that we can build up the of the resulting predicate based on the stem in a systematic fashion. It would need to have well-defined core units – the fundamental building blocks – and instructions for how these blocks fit together.

Two theories of semantics that aim to build meaning from basic primitive building blocks are Natural Semantic Metalanguage (Wierzbicka 1996) and Conceptual Semantics (Jackendoff 2002; Jackendoff 2010). However the nature of the building blocks differ in significant ways. While Jackendoff and Wierzbicka both express the innateness of the concepts or primitives in their lexical descriptions, only in the theory Wierzbicka expounds that these primitives, which are indefinable within the theory, have a special status — these primitives exist in all languages. Although these indefinables may exist in all languages, their interpretation and usage are culturally

mistake (*X* made a mistake)
something bad happened
because *X* did something
X didn't want it to happen
X wanted something else to happen
X thought that something else would happen

Figure 4.12: *mistake* in Natural Semantic Metalanguage

bound (Wierzbicka 1996:14-15). These primitives are 'the shared core of all natural languages (Wierzbicka 1996:22), and it is this core lexicon that provides the 'metalanguage for the description and comparison of meanings' (Wierzbicka 1996:23). For Jackendoff the primitives are the innate concepts in the mind, and therefore exist in all minds of humans beings, but determining which concepts should be added in the Conceptual Structure inventory of primitives is not as straight-forward in comparison to the well-bounded criterion of Natural Semantic Metalanguage.

The 'grammar' defined or rules for composing the building blocks for Natural Semantic Metalanguage lends itself well to be adapted for the task because it is simple. However, in its simplicity it can create more verbose semantic descriptions. For example, Wierzbicka (1996:280) proposed Figure 4.12, as the NSM representation for the single word *mistake*.

Although the notion of Natural Semantic Metalanguage, which consists those shared primitives of the world's languages, is a very romantic ideal, the restriction it poses for us in our goal would render our description of verbs a little unweildy, given that thus far little more than sixty primitives have been established (Goddard and Peeters 2006). We borrow the primitives from the Natural Semantic Metalanguage inventory, but for ease of representation we use the grammar defined for Jackendoff's (2010) CS. We outline the primitives we employ in describing affixed predicates relative to their stems in Section 4.4.4.

4.4.4 Primitives for Analysis

This section describes the basic units used in the decomposition of the meN+STEM and the meN+STEM+KAN verbs. We describe the building blocks we use in the scaffolding around the stem and its arguments to capture semantics changes between these two verb types.

The primes described here are mostly borrowed from Natural Semantic Metalanguage (Wierzbicka 1996; Goddard and Peeters 2006), but the way they are combined together is much closer to Jackendoff's Conceptual Structure (Jackendoff 2002;

Jackendoff 2010), because the 'grammar' of CS is more suitable for our purpose than NSM, in that it is more concise. We do not employ CS to its fullest because we do not aim to break down the description of each event fully, but only in relation to the stem, which describes the action, and the participants, or arguments, involved.

There are four kinds of events, which we adopt as part of our primitive lexicon in describing predicates relative to their stems. The three events **DO**, **HAPPEN**, **GO**, and **CAUSE** are shown in Figure 4.13 with their sub-types.

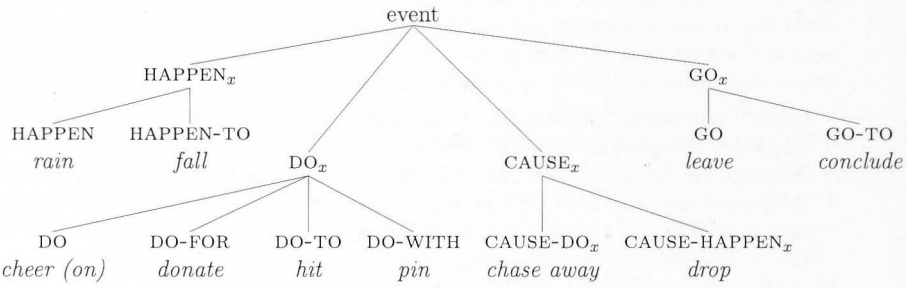


Figure 4.13: Basic primitive lexicon

A summary of the primitives we use in our task are as follows:

DO describes a single event that is triggered by volitional action, and as described by Wierzbicka (1996:122) opens up an 'agent' slot. However the presence of a 'patient' slot is less clear with this prime, and Wierzbicka suggests that an optional patient slot be allowed. This, according to Wierzbicka (1996:122–3), eliminates the need for separate DO and DO-TO (DO$_{to}$) predicates, with the latter allowing a patient slot. She also includes DO-WITH (DO$_{with}$), which allows an instrument argument, as the same predicate because these are simply DO with variations on valency options. We follow this recommendation, and consider these types of predicates as being the same event type, which we have labelled DO$_x$ in Figure 4.13, and include DO-FOR (with a benefactive argument) in this family of events.

HAPPEN events have no 'agent', but have a 'patient' or some experiencer in the HAPPEN-TO (HAPPEN$_{to}$) variant. Events in the HAPPEN$_x$ family describe single events with no deliberate or volitional action.

CAUSE has two sub-types: CAUSE-DO and CAUSE-HAPPEN, which describe their sub-event types. The CAUSE primitive is not in the Natural Semantic Meta-

language inventory, but instead there is the linker BECAUSE to express the notion of causality. We employ the Conceptual Structure style CAUSE to describe an event that is composed of subevents, unlike Wierzbicka's (1996) clause linking BECAUSE mainly for ease of representation.

GO is from CS not NSM, encodes not just physical paths and motion, but can encode change of state or events that result in a change of state.

Other primitives include negation and concepts introduced by prepositions such as NOT, TO, WITH, and FOR. In addition we include two states from the NSM inventory. BE simply describes a state, but FEEL is a state that has an element of affectedness on the experiencer. These states are normally employed with adjective stems. In the NSM taxonomy, the latter is classified under 'mental predicates'. As for noun stems, Jackendoff (2002:35) demonstrates that the semantics of denominal verbs are inherently idiosyncratic and that the interpretation of denominalised is often conventionalised and not entirely predictable. This makes our task in producing *semantic scaffolding* around the stem to be even more difficult for nouns. For this reason, we introduce 6 primes that can form a verbal unit with the nominal. These primes are PERFORM, ACHIEVE, MAKE, HAVE, CREATE, and BECOME. The predicate BECOME is part of the theory of Conceptual Structure, and is employed by Kroeger (2007) in describing the causative *-kan*. The other predicates were discovered as part of the process of mapping out and grouping syntactically-like verbs, and it is beyond the scope of this study to try to prove their cross-linguistic translatability to be incorporated into NSM.

We have not taken wholly from NSM, for example GO and CAUSE are from CS, and we do not use the NSM primitive BECAUSE to express causation, because it cannot be represented in CS syntax, and we also found that with the limited inventory for NSM, we had to introduce concepts to build up predicates with noun stems.

4.4.5 100 Verbs

From the method set out in the previous sections we arrive at 26 Types for the 100 verbs in our investigation: 8 verb, 13 noun, and 5 adjective types. If a stem is clustered into a type then they all alternate with respect to *-kan* in the same way.

The manually induced groups of stems that represents their Types are shown in Table 4.4, with the verb types listed first, followed by nouns, then adjectives. As we had noted in Section 4.4.1, we adhere faithfully to the categorie published in *Kamus Besar Bahasa Indonesia* (KBBI – 'The Big Indonesian Language Dictionary') (Sugiono 2008). In particular, we label a word with the part-of-speech category of the first sense listed in the dictionary entry.[9]

[9]In the KBBI, the entry *ajar* "lesson/learn" lists the nominal sense first and as such we list this

MORPHOLOGY	VERB FRAME	DECOMPOSITION
Adj Type 4: *kecewa* "disappointed", *leceh* "worthless", *remeh* "unimportant", *teguh* "strong", *jengkel* "annoyed"		
MEN+A_4	–	–
MEN+A_4+KAN	<NP_a >	CAUSE(NP_a, [FEEL(NP_a, A_4)])
Adj Type 5: *lunak* "soft", *lanjut* "protracted"		
MEN+A_5	<NP_a >	BECOME(NP_a, A_5)
MEN+A_5+KAN	<NP_a >	CAUSE(NP_a, [BE([NP_a] [A_1])])
Verb Type 3: *dengar* "hear", *kenang* "think of"		
MEN+V_3	<NP_a, NP_b >	$HAPPEN_{to}$(NP_b, [V_3 TO_{happen}[NP_a])])
MEN+V_3+KAN	<NP_a, NP_b >	DO_{to}(NP_a, [V_3 TO_{do}([NP_b])])
Verb Type 4: *hidup* "be alive", *jatuh* "fall", *mati* "die", *tewas* "perish", *pusing* "to concern oneself", *minggir* "put aside", *masuk* "enter", *hadir* "be present", *lulus* "go through"		
MEN+V_4		
MEN+V_4+KAN	<NP_a, NP_b >	CAUSE(NP_a, [$HAPPEN_{to}$([V_4 TO_{happen}([NP_b])])])

Table 4.3: Verb Frames and Semantic Decomposition: examples of discovered adjective and verb types

An example of each of syntactico-semantic mappings that we found in order to group these stems are shown in Table 4.3, with the full list shown in Appendix C. This table show an example from verbs, and adjectives.

Stems that fall under Type N_2, like many of the other noun types, allow for an unexpressed object in the verb frame for MEN+stemN2. The problem with these kinds of verbs is whether to analyse them as true transitives with unexpressed objects, or as ambitransitive sentences. In addition there is also the issue of representation: one could argue that for the former case, at the f-structure, they would have some object that is not expressed in the c-strucutre. For the latter case, these could be represented as two different lexemes - one that requires an object and one that does not.

Once defined in the lexicon, these types we have discovered can restrict how stems behave when affixed with *-kan*. However, there is one issue, and that is we only know how to apply this categorisation to 100 stems. Our goal would now be how to learn these *-kan* Types automatically. However, in the discovery verbs classes that behave syntactically in the same way, these are often arrived at through subcategorisation features, as seen in Section 2.6. This unfortunately would require a large scale parser

item as a noun, even though this word is most commonly used as a verb.

1.	V_1	*acuh* "to heed", *terjemah* "translate", *mandi* "bathe"
2.	V_2	*bawa* "carry", *beri* "give"
3.	V_3	*dengar* "hear", *kenang* "think of"
4.	V_4	*hidup* "be alive", *jatuh* "fall", *mati* "die", *tewas* "perish", *pusing* "to concern oneself", *minggir* "put aside", *masuk* "enter", *hadir* "be present", *lulus* "go through"
5.	V_5	*serah* "surrender", *singkir* "get out of way", *susup* "duck down"
6.	V_6	*bangun* "form/take shape", *pecah* "be broken"
7.	V_7	*paksa* "force", and also *buat* "make/do"
8.	V_8	*timpa* "hit", *baca* "read"
9.	N_1	*administrasi* "administration", *instalasi* "installation", *legalisasi* "legalisation", *nasionalisasi* "nationalisation", *ikat* "tie", *pukul* "blow/strike", *sewa* "hire"
10.	N_2	*ajar* "lesson"
11.	N_3	*gambar* "picture"
12.	N_4	*aplikasi* "application", *ekspresi* "expression", *kerja* "activity/work"
13.	N_5	*belanja* "expenses", *gelembung* "bubble", *buku* "book", *publikasi* "publication", *radiasi* "radiation", *kumandang* "echo"
14.	N_6	*darat* "land", *didih* "boiling"
15.	N_7	*hipotesis* "hypothesis", *titah* "command", *mimpi* "dream", *pikir* "idea", *tanya* "question"
16.	N_8	*asumsi* "assumption", *umpama* "example", *wakil* "proxy", *lokasi* "location"
17.	N_9	*paten* "patent" *tempat* "place" *tumpu* "foothold" *letak* "position" *penjara* "jail" *rumah* "hou
18.	N_{10}	*injeksi* "injection", *kait* "hook", *analogi* "analogy" *maklumat* "declaration"
19.	N_{11}	*sesal* "regret", *susu* "milk"
20.	N_{12}	*janji* "promise", *cerita* "news"
21.	N_{13}	*sarang* "web", *percik* "stain", *mula* "start", *kerja* "work"
22.	A_1	*abadi* "eternal", *asing* "separated", *cemar* "dirty", *cerdas* "intelligent", *goyah* "unstable", *haram* "prohibited", *murni* "pure", *mutakhir* "recent/up-to-date", *padu* "compact/solid", *populer* "popular", *salah* "wrong", *subur* "fruitful", *terang* "clear"
23.	A_2	*biasa* "ordinary/common", *unggul* "excellent/ahead", *berani* "audacious"
24.	A_3	*cengang* "amazed", *takjub* "surprised"
25.	A_4	*kecewa* "disappointed", *leceh* "worthless", *remeh* "unimportant", *teguh* "strong", *jengkel* "annoyed"
26.	A_5	*lunak* "soft", *lanjut* "protracted"

Table 4.4: All types

for Indonesian, which sadly we do not have, or have access to.

Although the lexical semantics of the stem were largely ignored in this task (and only changes to semantics with respect to *-kan* were noted), we found pockets of synonyms within the Types defined. For example, *mate* "die" and *tewas* "perish" both belong to Verb Type 4, *singkir* "get out of way" and *susup* "duck down" both belong to Verb Type 5. Within the noun types, *buku* "book" and *publikasi* "publication" have been categorised within Noun Type 5, and *tempat* "place" and *letak* "position" are both grouped into Noun Type 9, and the adjectives *leceh* "worthless" and *remeh* "unimportant" are both in Adjective Type 4.

This outcome should come as no surprise because as shown by Levin (1993), there

is a tight connection between lexical semantics and syntactic structure. Although, much of the work done on grouping syntactically-like verbs together have used sub-categorisation features, as mentioned above, we aim to see, given the tight connection between lexical semantics and syntactic behaviour, if we can employ computational method that models semantics in order to arrive at groups of verbs that have the same syntactic profile.

stem (S)	S	meN+S	meN+S+kan
*acuh** "heed"	30	0	14
baca "read"	528	1328	47
bangun "get up"	603	2876	61
bawa "bring"	149	4482	654
beri "give"	170	3384	7665
buat "make/do"	750	12651	84
dengar "hear"	152	1364	357
hadir "be present"	1123	0	245
hidup "live/be alive"	9687	0	297
jatuh "fall"	3788	0	452
kenang "think of"	22	404	8
lulus "go through"	1661	0	60
mandi "bathe"	681	0	34
mati "enter"	4120	0	393
masuk "die"	7557	0	939
*minggir** "put aside"	42	0	5
paksa "break"	375	812	119
pecah "force"	581	188	549
*pusing** "be concerned"	87	0	8
serah "surrender"	86	1047	9
singkir "get out of way"	0	62	348
susup "duck down"	0	126	8
terjemah "translate"	21	0	270
tewas "perish"	2699	0	684
timpa "hit"	0	248	16

Table 4.5: Frequency of occurence in Wikipedia for verb stems.

4.5 Discussion

One possible drawback of choosing the 100 stems at random to obtain our Types is that this method may not arrive at a representative sample of lexical variation that will give a comprehensive representation of syntactic behaviours, as often a small number of relatively high frequency lexemes tend to exhibit the most varied and interesting behaviour.

Table 4.5 shows the frequency of usage for the verbs stems in Wikipedia. The first figure in each line shows the frequency of the verb without any affixes, the second figure is for the number of times the stem appears in the morphological context MEN+STEM, and the last is for MEN+STEM+KAN.

As can be seen from Table 4.5, most of the verbs stems chosen at random are relatively high frequency verbs, with the exception of those starred '*'.

In this chapter we have detailed the implementation of aspects of deviant morphology in Indonesian. In particular, we have shown that although *-kan* is varied, if its application is not restricted then this will lead to overgeneration of the deep grammar. We mitigate this by defining types that constrain the behaviour of *-kan* according to its stem, and our task now is to be able to automatically learn these types, in order to expedite lexicon development and aid the lexicographer.

However, there are many assumptions that we make in defining these types and on embarking on the task to automatically acquire verbal information, such as syntactic alternations. We take it for granted that the definitions of our types are partitioned according to part-of-speech, and the fact that the definition of our rules in the deep grammar also use notions of word classes, despite this feature of Indonesian being in dispute (Gil 1994; Gil 2001; Gil 2010; Yoder 2010). Before we embark on conducting lexical acquisition on verbs, and relying on our word class partitioned types, we first performed an experiment that will enable us to rely on theses concepts as being valid in Indonesian.

In the following chapters in Part III, Chapters 5 and 6, we first perform an experiment to determine whether word classes can be established or not in Indonesian, and then we embark on learning syntactic alternations by exploiting distributional semantic methods.

Part III

Application of Deep Lexical Acquisition

Chapter 5

Investigating Indonesian Word Classes

5.1 Introduction

Word classes in many Austronesian languages have been proven to be difficult to define, for example, in seeking an adjectival class in Tukang Besi (Donohue 2010), and for distinguishing categories for open class words in Tagalog (Foley 1998; Foley 2008; Kaufman 2009). Included in this list of languages, where defining word classes is proven to not be straightforward, is Indonesian, as it spoken in certain parts of the country, such as Riau Indonesian (Gil 2001) and Jakartan Indonesian, as acquired by children (Gil 2010).

It has been claimed that from a morpho-syntactic perspective major word classes in Indonesian are indistinguishable. That is:

> [A] salient grammatical feature of many Austronesian languages is the similar or identical morphosyntactic behaviour exhibited by expressions denoting things (eg. boy, bird, helicopter), and expressions denoting activities (eg. walked, took, gave). (Gil 2001)

This chapter looks at the claim that in Indonesian, all open class lexical items are indistinguishable with respect to their parts of speech (Gil 2001; Gil 2009; Gil 2010). That is to say there exists only one open class category, which means that verbs are indistinguishable from adjectives, adverbs, and most importantly nouns. We test this claim for the language as it is written by a variety of Indonesian speakers using empirical methods commonly used in part-of-speech induction.

We design an experiment that utilises only morphological patterns in ascertaining word classes. This is a complementary study to Yoder's (2010) who show that at the phrase structure level, there are lexical insertion points that are associated with word classes. However, it is shown that in languages, there can be a mismatch between the

categories at the syntactic level and the word class categories at the morphological level. For example, Himmelmann (2008) shows for Tagalog that there are syntactically defined classes, but for word formation all stems belong in one open word class. This task we define is not unlike the part-of-speech (POS) induction tasks we discuss in Section 2.6.2. However, these tasks primarily use collocation features in determing classes and commonly do not rely on morphological features, even though they have been found to boost performance (Christodoulopoulos *et al.* 2010).

The morphological features we employ are generated automatically from the morphological analyser discussed in Section 3.2.2.[1]

In this study, we find that once the distribution of the data points in our experiments match the distribution of the text from which we gather our data, we obtain significant results that show a distinction between the class of nouns and the class of verbs in Indonesian. Furthermore it shows promise that the labelling of word classes may be achieved only with morphological features. Morphological features are often secondary features or are accompanied with collocational features in part-of-speech induction tasks. From this study we see that the utilisation of solely morphological features for this kind of task is viable, and there shows promise in the application of out-of-vocabulary items.

5.1.1 Motivation

The notion of word classes, such as nouns, verbs and adjectives, is fundamental in both linguistics and computational linguistics. Word classes are the basis for the labels in part-of-speech tagging, and also the building blocks for parsing. In grammar engineering, they are the primitives upon which context-free grammar rules are written. Part-of-speech tags are also widely utilised in natural language processing because they abstract away from the surface word and are therefore ways of alleviating the problem of data sparseness. In linguistics, they are considered to be the categories that shape the organisation of the language. These categories frequently do not align from language to language; what is expressed as a verb in one language may be expressed as an adjective or noun in another, but one claim that remains despite these variations is that the categories *noun* and *verb* exist in all languages (Croft 2003).

The motivation for this study is two-fold: In the field of linguistics it has been claimed that the noun–verb distinction is universal (Croft 2003), however this claim has been questioned for Indonesian (Gil 2010). We use empirical methods to ascertain whether this distinction holds true for this language. Second, in the natural language processing community there may be a strong reluctance to do away with such fundamental concepts as noun and verb, even if it were linguistically motivated. Computational linguists and grammar engineers working in Indonesian may be disin-

[1]However we modify this tool for the task at hand as described in Section 5.3.2.

clined to rewrite grammars on one open class category, rather than take advantage of the categorial distinctions nouns and verbs make in a language. However, it may be advantageous to know whether nouns and verbs are encoded for computational convenience or because the language is organised as such. In addition these experiments are a litmus test to see if morphological features alone can suffice in determining parts of speech. The methods employed in this study offer an avenue to explore the part-of-speech prediction of out-of-vocabulary items (OOV) in Indonesian text processing.

5.1.2 Assumptions

One assumption we make in this study is that stylistic variation and variation based on genre (such as spoken versus written) do not have differences as fundamental as dissolving the distinction between major word classes completely. That is, if the language as it is written should be analysed and described as having certain word classes, these classes should also exist, although not necessarily without difference, but exist nonetheless, in its spoken form. We also assume this for dialectal variation, and we assume that given two varieties of a language, they could not be syntactically so divergent that the entire word class system is at extreme ends, without mutual intelligibility being an issue.

However, this assumption is of course loaded with controversy, which we will not delve into – it is difficult to determine the fine line between what constitutes a language in its own right, and whether its a colloquial or dialectal variation. There are certainly extremely divergent dialectal varieties spoken throughout Indonesia, and even the difference between written Stadard Indonesian, and the spoken variety is astoundingly different, making it appear as isolating a variety as Riau Indonesian. Yoder (2010) indeed shows that many word order variations shown in Riau Indonesian, are indeed common focus strategies in Standard Indonesian.

The rest of the chapter is laid out as follows. In Section 5.2 we briefly discuss for Indonesian some of the problems that arise from the linguistic method of determining word classes, which we introduced in Section 2.4. We look at the formal properties of Indonesian, and we give examples to show how the distinction between nouns and verbs can be difficult to determine. For this reason, we take a computational linguistic approach, specifically we use unsupervised clustering (see Section 2.6). In the next two sections, Sections 5.3 and 5.4, we describe the data, tools and method for our experiments, followed by the results in Section 5.5. Finally, we discuss our findings and the impact and contribution to both linguistics and computational linguistics for Indonesian in the final sections 5.6 and 5.7.

5.2 Word Classes in Indonesian

We briefly restate the linguistic method we adopt in the discovery of word classes from Section 2.4 summarised in Table 5.1. This linguistic method is rather formal relying on the combinatorics of the form. At the clausal or phrasal level, we look at the syntagmatic possibilities of the units within the phrase or clause. When looking at the word level, we look at how each of the morphological components combine. The heuristics for this *combinatorics* approach are outlined by Evans (2000), which takes into consideration semantic properties as a way of labelling these classes rather than determining them.

However, when form and function do not generally coincide for the word classes assumed in a language, then this is exceptional. Furthermore when groups of words that traditionally serve different functions, for example ones that refer to an entity and ones that attribute a property, do not show distinct formal differences, then this may be a case of these two groups of words not being linguistically distinct. This has been the claim for Indonesian, and we want to ascertain whether the descriptions of the language involving distinct categories such as 'noun', 'verb' and 'adjective' in Indonesian as per Muhadjir (1981); Sneddon (1996); Mintz (2002); Sneddon *et al.* (2010) use these labels as a mere convenience with the full cognizance that these are employed as semantic labels and are not strictly morphosyntactic word classes.

i. Equivalent combinatorics
"Members of what are claimed to be merged classes should have identical distributions in terms of both morphological and syntactic categories."

ii. Compositionality
"Any semantic differences between the uses of a putative 'fluid' lexeme in two syntactic positions (say argument and predicate) must be attributable to the function of that position."

iii. Bidirectionality
"[T]o establish that there is just a single word class, it is not enough for Xs to be usable as Ys without modification: it must also be the case that Ys are usable as Xs.

Figure 5.1: Criteria for determining word classes

A case such as Indonesian is exactly the kind of instance where we can employ Evans and Osada's (2005) criteria, seen in Figure 5.1, for testing if word classes were justifiably merged by Gil (1994, 2010). However, it is exactly these kinds of criteria that Indonesian can satisfy. It can indeed be seen that the combinatorics of lexical items could lead one to analyse the language as having one open class category.

(5.1) *Ia lari.*
(S)he run
"(S)he runs."

(5.2) *Lari menyehatkan.*
run cause.to.be.healthy
"Running is healthy."

Looking at the *Equivalent combinatorics* criterion, in the following examples it can be seen in the example above that *lari* can occupy either the *subject* and the *predicate* position.

The examples below employ the stems[2] *bunyi* 'sound' and *bangun* 'to wake up', and show how words that can be traditionally thought of as a noun and a verb, respectively, can occur in the same morphological environment, by combining with the same morphological affixes:

(5.3) *membunyikan*
mem-bunyi-kan
AV-sound-KAN
"make X make a sound (instrument)"

(5.4) *membangunkan*
mem-bangun-kan
AV-wake.up-KAN
"make X wake up (someone)"

In this instance, we see that the affixes seem to not discriminate which stems they attach to. Take first the issue of *Compositionality* from Section 5.2, which states that given a word in a position, we should be able to predict its semantics. We see that in Sentences (5.1) and (5.2), the predicative or referential function of the word *lari* 'run' is simply determined by its position in the clause, unlike English that requires us to employ derivational morphology.[3]

In Examples (5.3) and (5.4), the semantic and morphosyntactic effects of the affixes are predictable: the suffix *-kan* forms a *causative* in both examples. The prefix *mem-* AV simply tells us that the agent is the subject, which would be true for both examples above.

Testing the criterion of *Bidirectionality* is not a trivial task for two reasons: (1) relying on grammaticality judgements can at times be difficult (Keller 2001); and (2) often the number of linguistics examples for such studies are rather small and would not be a representative sample of the language. The *Bidirectionality* criterion states that those classes traditionally labelled as nouns, for example, must be able to behave as all other parts of speech, and vice versa, if all other parts-of-speech are to be merged. We have an example of what would be traditionally analysed as a verb in a syntactic slot normally reserved for nouns in Sentence (5.2). However, we would also have to find possibilities of traditional nouns in verbal positions, without the need for morphological affixation to license its usage in that position.

[2]We use the term 'stem' to mean a word with no overt affixation, rather than a word-class-neutral form (as defined in (Bauer and Hernandez 2005:14))

[3]In English we can say: *I run.* but not: **Run is healthy.*

Rather than rely on grammaticality judgements and a small subset of open class words for testing, we rely on how the language is used by the Indonesian speaking population in the form of publicly available web data. This data represents a large number speakers and has in excess of 26 millions tokens.

5.3 Experimental Set up

This section outlines how we developed our data in experiments, obtained our evaluation data, and modified existing tools for the task in order to develop our morphological features.

The text we use for this study is the Indonesian Wikipedia because not only is it a large source of text, but also because the data is produced and curated by many authors; it is representative of the way the language is used throughout the Internet-connected areas of Indonesia, and Indonesian speakers throughout the world. We gathered approximately 26 million Indonesian tokens from Wikipedia articles and cleaned the existing mark-up. After tokenising, the data was sorted, and tokens counted (see Section 3.3.1 for details on how we prepared the Wikipedia data).

We then ran our morphological analyser over all tokens that occurred 5 or more times, mainly as a way of eliminating spelling and other errors. In actuality this is only around 17% of all word forms found, with a long tail of duplicate and singleton occurrences. Minimising errors, such as spelling errors, is important for the different data representations we employ, namely the *type* data described in Section 5.4.1. The type data, as opposed to the token data, indicate whether there is an attested or non-attested word form found in the corpus, and a typographical error in the corpus would result in an erroneous positive feature.

5.3.1 Stem lexicon

The method we employ in discovering the word class clusters is unsupervised, however, our main aim is to determine whether the way in which the stems are used suggests a clustering of nouns and verbs as separate categories. Hence, we do not employ an intrinsic evaluation of our discovered clusters, but instead compare them with data that we consider to be correct, insofar as their being entered into the lexicon of the morphological analyser in consultation with the *Kamus Besar Bahasa Indonesia* 'The Big Indonesian Language Dictionary' (Sugiono 2008) (see Chapter 4 for more on the morphological analyser).

This stem lexicon from the morphological analyser, with their parts of speech assigned to them, is used as the gold standard for our evaluation. For our experiments we altered our morphological analyser so that all word classes were treated as though they belonged to one large class allowing all stems to be treated equally.

The stem lexicon we derive from our machine readable grammar is biased towards

Part-of-Speech	Count
Noun	8,096
Verb	821
Other –	1,770
TOTAL	10,687

Table 5.1: Part-of-speech distribution in stem lexicon for morphological analyser.

Prefix	Actor Voice; Passive Voice; Causative; PassiveTer/ter; ber; OrdKe/ke; pe/peN; se
Circumfix	ke_an; per_an; pe/peN_an
Suffix	an; i; kan; wi
Clitics	ku; mu; nya
Other	Reduplication

Figure 5.2: Types of affixes from the morphological analyser

nouns, with almost ten times as many nouns as there are verbs, as can be seen in Table 5.1, which affects our initial experimental set-up. This proportion of nouns to verbs is artificial and does not necessarily reflect the proportion in naturally occurring text; it simply happpens to be the ratio in the lexicon for the morphological analyser.

However, we overcome this bias by subsampling and better representing, in our experimental data the proportions found in naturally occurring text, as seen in Section 5.5.2. The class labelled "Other" consists of all stems not marked as a noun or verb, and includes adjectives, pronoun, prepositions, numbers, and determiners.

5.3.2 Class Independent Morphological Analyser

The morphological analyser, built using XFST (Beesley and Karttunen 2003),[4] is defined with 10,687 stems in the lexicon, as well as the affixes outlined in Figure 5.2. The initial distribution of the stem/word classes in the lexicon is outlined in Table 5.1.

The morphological analyser was initially defined such that the class that a stem belonged to restricted how it combined with certain affixes in order to derive another word class. For this study, we modified this to relax any word class restrictions.

[4]As noted in Section 3.1, we employ XLE and XFST for developing grammar engineering resources because these are the tools used within ParGram.

For example, the prefix *ke* in Indonesian is affixed to a numeral in order to create ordinals, as in *ke+tiga* 'ord+three' would give us *ketiga* 'third'. In the morphological analyser the affix defined as ORDKE can only combine with stems that are defined in the lexicon as NUM (numeral). Another restriction is on the circumfix *peN+an*, which can affix to a verb stem in order to produce a noun, such as *pem+bakar+an* (peN+burn+an) meaning 'the process of incinerating'.

However, we relaxed any restrictions on stems so that all stems are treated the same in that all stems were categorised as once kind of word class, which we call 'LEX' to mean *lexical item*. This leads to uninhibited over-generation, but it also enables the possibility of analysing all classes of stems in a uniform manner.

5.4 Method

5.4.1 Feature Engineering

We have two kinds of experiments based on the values of our features: token and type features. Token features take into account the number of occurrences of a morphological pattern, and the type features have a positive value if there is an occurence of a particular morphological pattern.

The morphological patterns are collated from the output of the morphological analyser. By morphological patterns we mean all of the combinations of affixes that are attached to each stem in the lexicon, and not simply each individual affix which is appended to each stem. The combination of these affixes make up each feature.

If for example we had the excerpt from Figure 5.3 as our corpus, and we were only interested in collecting data for the lexemes *eat* and *pancake*, then the features for this collection would be as shown in Figure 5.4. The **a.** rows indicate the token values and the **b.** rows are the binary (type) values.

As seen in Table 5.4, we find a relatively small number of morphological patterns after running the morphological analyser over Wikipedia. The patterns can be made up from a combination of affixes in Figure 5.2). However, the combining of these affixes are restricted to what's possible in the language, even though these combination of affixes can apply to any type of stem. For example the suffix *-i* cannot combine with the suffix *-kan*, and this is reflected in the morphological analyser.

Table 5.2 gives an indication of what our features look like with actual token counts, and Table 5.3 shows our type data would look like.

5.4.2 Clustering

For our clustering experiments we have partitioned the data in various ways depending on the design of the experiments. Table 5.4 shows the proportion of instances in the data set that are nouns ("N"), verbs ("V"), or neither nouns nor verbs ("O")

> *We were all sitting around the big kitchen table. It was Saturday morning.* ***Pancake*** *morning. Mom was squeezing oranges for juice. Henry and I were betting on how many* ***pancakes*** *we each could* ***eat****. And Grandpa was doing the flipping.*
>
> *Seconds later, something flew through the air headed toward the kitchen ceiling...*
>
> *And landed right on Henry.*
>
> *After we reaslied that the flying object was only a* ***pancake****, we all laughed, even Grandpa. Breakfast continued quite uneventfully. All the other pancakes landed in the pan. And all of them were* ***eaten****, even the ones that landed on Henry.*

Figure 5.3: Excerpt from 'Cloudy with a chance of meatballs' by Judi Barrett (1982)

Morphological Patterns:		STEM	STEM+en	un+STEM+s
eat	a.	1	1	0
	b.	1	1	0
pancake	a.	2	0	2
	b.	1	0	1

Figure 5.4: Features extracted for *pancake*, *eat*

for each of the experiments that we run. The label "NoRR" refers to data that has no *morphological reduplication*, which for these experiments refers to the doubling of a stem (and in combination with other allowable affixes with reduplication). We omitted this feature for some of our experiments because it is claimed that Indonesian only exhibits derivational morphology, except for reduplication within an LFG framework (Musgrave 2001).[5] Reduplication applies to a number of word classes, but given they are inflectional, they should have meanings idiosyncratic to their respective category. We preempted that this may affect the automatic assignment of instances to clusters and we decided to create another subset of experiments we call NoRR, which omits all reduplicated forms.

We also ran an experiment that had only instances that were nouns and verbs because stems labelled "O" (other) belonged to very disparate classes including pro-

[5]However, in the Minimalist framework voice marking is considered the inflectional head of in the projection immediately above the VP node (Son and Cole 2008:137)

instance	stem	stem+an	per_an+stem	redup[stem]	me+stem+kan	me+stem+kan+nya	...
konfirmasi	81	0	0	0	0	0	...
tangkal	9	0	0	0	0	0	...
parkir	607	16	0	0	6	0	...
main	978	415	81	42	1972	69	...
diam	413	8	0	426	8	0	...
makan	2584	5607	0	8	0	0	...
sikat	67	0	0	0	0	0	...
bandung	5434	0	0	0	0	0	...
esai	286	0	0	17	0	0	...
kitab	4417	0	0	444	0	0	...
...	...	...	...	...	...	...	...

Table 5.2: Morphological patterns for Indonesian: Token data

instance	stem	stem+an	per_an+stem	redup[stem]	me+stem+kan	me+stem+kan+nya	...
konfirmasi	1	0	0	0	0	0	...
tangkal	1	0	0	0	0	0	...
parkir	1	1	0	0	1	0	...
main	1	1	1	1	1	1	...
diam	1	1	0	1	1	0	...
makan	1	1	0	1	0	0	...
sikat	1	0	0	0	0	0	...
bandung	1	0	0	0	0	0	...
esai	1	0	0	1	0	0	...
kitab	1	0	0	1	0	0	...
...	...	...	...	...	...	...	...

Table 5.3: Type data

nouns, determiners, adjectives, and numerals.

Rather than applying a hard clustering algorithm that assigns each data point to its closest centroid, such a K-means, we decided to employ a soft probabilistic clustering algorithm, namely the Estimation Maximisation (EM) algorithm. The reason we opt for a soft clustering algorithm is to reflect the fact that stems can, and do, belong to multiple classes.[6]

We employ the EM implementation in the Weka package[7] (Witten and Frank

[6]Although we have a probabilistic assigment of stems to classes in our experiments, in our evaluation we have a categorical gold standard because there is no resource that states that a stem is 40% adjective and 60% verb, and it would be problematic to create such a resource. Instead, we get partial 'credit' for the probabilistic assignment the soft clustering assigns. Also, note that these experiments employing the EM algorithm were conducted before the experiments in Chapter 6 that use HDP, which could equally work as well for these experiments.

[7]`http://www.cs.waikato.ac.nz/ml/weka`

Data	# Features	N	V	O	Total
All POS	259	8,096	821	1,770	10,687
Noun & Verbs	250	8,096	821	-	8,917
All POS, NoRR	208	8,096	821	1,770	10,687
Nouns and Verbs, NoRR	200	8,096	821	-	8,917

Table 5.4: Part-of-speech distribution for the different experiments.

2005), maintaining the default parameters for the maximum number of iterations ($I = 100$), the minimum allowable standard deviation ($1e - 6$), and the number of random seeds initially selected ($S = 100$). We only changed the number of clusters to be found from $N = -1$ (where the number of clusters is determined via the setting of N that maximises log likelihood) to $N = 2$ (learn 2 clusters) for a subset of our experiments.

5.4.3 Experimental Procedure

We perform 2 groups of experiments which we call the 'All Clustering' and the 'Subsampling'. The 'All Clustering' experiments automatically clusters all the stems defined in the morphological analyser. There are two sub-experiments in 'All Clustering' that involve all parts of speech (ALL POS), with the second with only nouns and verbs (NOUNS AND VERBS). The ALL POS experiments are evaluated on a three way split of "N", "V", and "O" (nouns, verbs, and other), while the second only involves "N" and "V". The ALL POS experiment comprises of 10,687 stems, while the NOUNS AND VERBS have 8,917 stems. This proportion of nouns to verbs is however unbalanced, as our stem lexicon is noun-heavy with 75.8% of the stems being nouns, and in the experiments involving only "N" and "V" there is a proportion 90.8% nouns-to-verbs, as shown previously in Table 5.4.

For the 'Subsampling' experiments, we aimed to reduce the bias towards nouns. However, rather than take a fifty-fifty split of "N" and "V", we wanted a data split that was representative of the split seen in the text from which we gathered our data points. In order to get this information, we hand analyse a small Wikipedia entry, the Indonesian *Linguistik komputasional* 'Computational Linguistics' stub article.[8] For each unique token in the article (except English words, words in the footer, the menu, and tab items not relevant to the document), we consulted the Indonesian government's official dictionary *Kamus Besar Bahasa Indonesia* 'The Big Indonesian

[8]Accessed May, 2011 at the URL: `http://id.wikipedia.org/wiki/Linguistik_komputasional`

Language Dictionary' (Sugiono 2008) for its word class.

We found that the proportion of verbs to nouns was around 35%, with 16 verbs and 30 nouns, some being both verb and noun, which were counted twice, once for each category. Equipped with this knowledge, we sub-sampled based on these proportions and re-ran the experiments. For these 'Subsampling' experiments also ran two types: one with a 650–350 mix of nouns to verbs, and another with a 1300–700 noun–verb combination.

5.4.4 Evaluation

As briefly discussed in Section 5.3.1, the way we evaluate the discovered clusters is by ascertaining how well they align with the categories assigned in stem lexicon originally devised for the morphological analyser described in Chapter 4, which distinguished part-of-speech categories. For experiments with the setting of $N = -1$ multiple clusters could be found, and in these cases we report the combination of clusters and assignment of classes that yielded the highest F-score.

In the evaluation of the induced clusters, we compare them against two baseline systems, namely 'Majority Class' and 'Random'. With the 'Majority Class' baseline we build a system that classifies all instances as "N", being the majority class; and with the 'Random' baseline we randomly assign each instance as noun, verb, or optionally other depending on the system we compare against.[9]

We calculate *precision*, *recall*, and *F-score*, as described in Section 3.3.6, to ascertain how well our induced classes, based on morphological features that are agnostic to stem classes, can reproduce the word class divisions defined on the KBBI. We also conduct significance testing using a non-parametric method called *random sampling* (Yeh 2000) with 10,000 iterations.

5.5 Results

In this section we report the clustering results from the 'All Clustering' and 'Subsampling' experiments.

The columns N-V-O and N-V in Tables 5.5 and 5.6 indicate the number of clusters found and how they were merged for evaluation. For example under N-V-O in Table 5.5, the *Token* row reads 1-1-2, which indiates that four clusters were induced with two of them combined to form the "O" class, one for the "N" class, and one cluster for "V" class. In Table 5.6, under N-V in the *Type* row, we see that 4 clusters were

[9]It is not uncommon for majority class baselines to be used in part-of-speech or word class induction research, for example Biemann (2006). However, our aim is not to provide a state-of-the-art POS induction system for Indonesian. This is unlike Wicaksono and Purwarianti's (2010) aim who employ a 'weaker' POS induction system as their baseline. In our case we want to ensure that nouns and verbs can be differentiated and employ methods from research on POS induction. Given this aim the baselines we have provided are fitting for the task.

Data Type	All POS $N = -1$			
	P	R	F	N-V-O
Token	.985	.588	.737	1-1-2
Type	.977	.579	.727	1-1-3
Random	.623	.341	.441	–
Majority Class	.756	.756	.756	–
Token (NoRR)	.989	.591	.740	1-1-4
Type (NoRR)	.940	.556	.699	1-1-1
Random	.616	.338	.436	–
Majority Class		.757		–

Table 5.5: "All Clustering" Results for all word classes (N-V-O).

induced with three combining to form the "N" class with the other being evaluated at the "V" class.

5.5.1 "All Clustering" Results

As can be seen in Table 5.5, the experiments over all parts of speech, fared rather poorly, with F-scores (F) falling below the (supervised) majority class baseline. Bear in mind, however, that our primary question is whether open word classes, and in particular nouns and verbs, can be distinguished in Indonesian. However, these experiments included nouns, verbs and a heterogeneous assortment of closed- and open-class words.

When we omit closed classes, and all other classes except nouns and verbs in our experiment, we observe that our precision (P), recall (R), and F-score (F) all exceed the majority class baseline, for both the Token and Type systems, as seen in Table 5.6. The highest F-score was for the system that included reduplication, achieving an F-score of .990 and .973 for the Token and Type systems.

The results we found most surprising were from the systems that had no morphological reduplication (NoRR). These were consistently outperformed by the systems that had reduplicated stems. We return to discuss this in Section 5.6.

In order to ascertain whether these results are significant, we employed a computationally-intensive randomised test called *randomised shuffling*, which makes no assumptions about the distribution of the data (Yeh 2000). At this stage, none of the differences in results over the Majority Class baseline are statistically significant ($p > 0.05$), although all of the differences over the random baseline are significant ($p < 0.001$).

Data Type	Nouns & Verbs						
	$N = 2$			$N = -1$			
	P	R	F	P	R	F	N-V
Token	.930	.931	.930	.990	.990	.990	1-1
Type	.914	.915	.915	.973	.973	.973	3-1
Random	.512	.512	.512	.512	.512	.512	–
Majority Class	.908	.908	.908	.908	.908	.908	–
Token (NoRR)	.942	.942	.942	.942	.942	.942	4-1
Type (NoRR)	.918	.918	.918	.971	.971	.971	3-1
Random	.501	.502	.502	.501	.502	.502	–
Majority Class		.908			.908		–

Table 5.6: "All Clustering" Results for N-V experiments.

5.5.2 "Subsampling" Results

The subsampling experiments are a better reflection of the relative occurrence of nouns and verbs in actual text. Even though we had attained positive results against the random baseline, we had initally anticipated attaining poor results in the design of the 'All Cluster' experiments due to the high proportion of nouns compared to verbs in the morphological lexicon with over 75% of the total lexicon being nouns with 8115 stems, and under 8% being verbs. If we only take into consideration nouns and verbs then the percentage of nouns is around 90%. The disproportionate number of noun stems in our experimental data was due to the high number of nouns in the lexicon, and not a reflection of naturally occuring text.

We ran two types of experiments: one with a 650–350 mix of nouns to verbs, and another with a 1300–700 noun–verb combination, as described in Section 5.4.3. We found that having a proportion of nouns and verbs that was more representative of the text from which we generated our data had positive results.

Firstly, all F-scores surpassed the 0.650 majority class baseline, shown in Tables 5.7 and 5.8. Secondly, almost all of these results were significant based on our random sampling method, as shown in Table 5.7 for $N = 2$ and Table 5.8 for $N = -1$. These tables also show that on the *Token* 650-350 experiments did not produce significant scores for both the regular and NoRR results. As with the previous 'All Cluster' experiments, all F-scores exceeded the random baseline.

The boldfaced figures in Table 5.8 highlight the results that exceed the majority-class baseline at a level of statistical significance ($p < 0.05$). As can be seen, in most cases there was a significant improvement over the baseline for both the smaller

Data	$N = 2$			
	650–350	p	**1300–700**	p
Token	.783	.184	**.788**	**.043**
−NoRR	.778	.452	**.803**	**.000**
Type	**.812**	**.000**	**.825**	**.000**
−NoRR	**.812**	**.001**	**.828**	**.000**
Random	.538		.529	
Majority Class	.650		.650	

Table 5.7: Results for "Subsampling" experiments

Data	$N = -1$					
	650–350	**n-v**	p	**1300–700**	**n-v**	p
Token	**.808**	3-1	**.000**	**.788**	1-1	**.044**
−NoRR	.780	2-1	.490	**.793**	3-1	**.003**
Type	**.822**	3-3	**.000**	**.848**	5-4	**.000**
−NoRR	**.824**	3-4	**.000**	**.849**	3-5	**.000**
Random	.538			.529		
Majority Class	.650			.650		

Table 5.8: Results for "Subsampling" experiments

and larger datasets. This provides strong weight to the claim that the noun–verb distinction is discernible for Indonesian. With the exception of Token $N - 1$, the results for the experiments with 2000 stems were slightly better than the 1000 stem experiments, and the Type experiments better predicted word classes than the Token experiments. Again, the systems that omitted reduplication (NoRR) did not perform as well as expected.

5.6 Discussion

5.6.1 Reduplication

The results we found most surprising were the data that had no morphological reduplication (NoRR). These did not convincingly outperform systems that had reduplicated stems. We had expected the NoRR systems to perform much better than those with reduplication because, as stated by Musgrave (2001) reduplication is the

only inflectional morphology in Indonesian. If this is the case, then we would expect reduplication in nouns and in verbs to not have shared features. That is, the process of reduplication, applied to nouns should impose different semantic interpretations when applied to verb. We expected both nouns and verbs to participate in reduplication readily, but these, we thought should represent different constructions across these word classes and these processes should be opaque to our system.

We had expected that reduplication in Indonesian was much like the suffix *-s* in English – the same form performs different functions when applied to different classes: NOUN-*s* forms a plural, and VERB-*s* is the third person singular form of the verb. Given that we assumed that form and function was not uniform over the word classes for Indonesian reduplication, we predicted that it should add noise to our system. However, the experiments showed that omitting reduplication did not have a big impact on the results.

Table 5.5 shows that reduplication primarily applies to nouns, and for this class it encodes a plural meaning, as seen in Example (5.5) for the noun *kursi* "chair".

(5.5)

a. *Saya membeli kursi*
1.sg AV+buy chair
"I bought chairs / a chair."

b. *Saya membeli kursi-kursi*
1.sg AV+buy chair-chair
"I bought chairs / various types of chairs."

However, this plural semantics for reduplication does not just apply to nouns. For the predicative adjective *sakit* "sick", in Example (5.6), when it is reduplicated can force a plural interpretation of the subject *anak* "child". This is also the case with the verb *mati* "to die".

(5.6) *sakit* vs. *sakit-sakit*

a. *Anaknya sakit*
child+3.POSS sick
"His/her child is sick."

b. *Anaknya sakit-sakit*
his/her.child sick-sick
"Each of the children were sick"
"The child was sick (on and off for a period)"

(5.7)

a. *Orang mati*
person die
"The person died"

b. *Orang mati-mati*
person die-die
"The people died"/"*The person was dying."

We see in Example (5.7b), the only interpretation available in this reduplicated form is the forcing of a plural subject. Typically reduplicated verbs are interpreted as having progressive aspect, as seen in *mempukul-pukul*.

(5.8)

a. *Dia memukul temannya*
3.sg AV+hit friend+3.POSS
"He/she hit his/her friend."

b. *Dia memukul-mukul temannya*
3.sg AV+hit-hit friend+3.POSS
"He/she is hitting his/her friend (also repeatedly)."

However, hit is a punctual action, and one could interpret *memukul-mukul* as the multiple application of the same action. Although it appears that reduplication across word classes may be semantically related after all and not that idiosyncratic, the fact that the NoRR results did not fare well is not necessarily evidence for this.

In fact, upon doing a post-hoc analysis to determine why we had got this unexpected result, we found that only approximately 10% of the stems in our experiment displayed any reduplication at all. We found approximately 23,500 unique word forms in our corpus, given the stems we had investigated. Of those unique forms only 7% were reduplicated forms. For the token counts, we had gathered almost 9 million tokens, however only over 1% of these were tokens with reduplication. These low figures, particularly the low token counts, may be the reason why the NoRR experiments had not made as much of an impact as we had initially thought they would.

5.6.2 Morphological Features in Determining Word Classes

One experimental question we had was whether using only morphological features would suffice in determining word classes, and whether syntactic features were a requirement for this kind of experiment. We also were curious to see if there were morphological patterns that defined a class. For each of our features in the 1300–700 type experiment, we collected the accumulated probabilities for each class and compared them to each other. We found that there were about 10 morphological

Features with Greatest Probability Density as a Proportion per Class			
Noun		Verb	
STEM+NYA	SE+STEM	STEM+PEN_AN	TER+STEM
REDUP[STEM]	STEM+I+NYA	Passive+STEM+KAN	Active+STEM+KAN+NYA
STEM		Active+STEM+KAN	Passive+STEM+I
BER+STEM		STEM+AN+NYA	STEM+KAN
REDUP[STEM]+NYA		STEM+AN	Active+STEM+I

Figure 5.5: Morphological patterns associated with nouns and verbs.

patterns that we could associate with verbs, and 7 with nouns. These are shown in Figure 5.5

With respect to our evaluation, we modeled multiple word class membership by allowing a probabilistic assignment of stems to classes, even though we have a strict evaluation of them, as belonging to only one class. This strict evaluation, which only assigns the primary sense (as it is found in the KBBI) to each word only puts our method at a disadvantage because we can only get partial credit for stems that do indeed linguistically fall in multiple classes because of the shortcomings of the gold standard data. Therefore a positive result in these experiments can be interpreted positive despite the initial disadvantage imposed by the method of evaluation.

The results suggest that, although Indonesian does not have true inflectional morphology, which is usually the morphological basis for determining word classes (see Section 2.4), there are associated morphological patterns for each word class. If we have a PASS or AV prefix, then we would expect a verbal stem. However, a *-nya*, which can be a third person possessive clitic or a definiteness marker, suggests we have a noun stem, unless there is also an AV prefix, when renders *-nya* as an argument, representing a third person patient.

5.6.3 stem+i+nya vs. stem+nya

The STEM+I+NYA may seem like a very strange morphological pattern for nominals, but this may have something to do with the '-i' suffixation rule for when a stem ends in 'i'. In Indonesian, vowel hiaitus is avoided at the right edge of the stem (Cohn and McCarthy 1998), and this is not usually reflected in the orthography when hiaitus is avoided by inserting an epenthetic glide, as shown in Figure 5.6.

However, Sneddon (1996:84) states that "when suffix *-i* occurs with a base ending in *i* the sounds merge into one", as shown in Figure 5.7.

Therefore, all stems that end with *i* optionally allow an *-i* suffix in the morpho-

bantu+an → bantu|wan "help" (nom)
jadi+an → jadi|yan "case"

Figure 5.6: A subfigure from Cohn and McCarthy (1998)

beri+i → beri

Figure 5.7: The verb beri

logical analyser.

Given that the morphological pattern STEM+NYA is a strong indicator the word class of the stem is a noun, we can assume that the pattern STEM+I+NYA accounts for the stems that end with *i*, that has been appended with an 'invisible' *-i* suffix.

5.6.4 Type vs Token

From a linguistic perspective, one result that we feared was that the experiments that took into consideration the number of times a morphological patterned appeared for each stem (token), would fare better than the type experiments. The implications of this would suggest that the means by which linguists determine word classes, by adding to their inventory of possible combinatorics, would not suffice. If a 'prevalence-based' approach was required, that suggests that linguists not only had to take into account the fact that a form is possible, but also how often it is likely to occur, which is not a standard approach in descriptive linguistics.

However, we see that the type experiments have the same score or do better than the token-count experiment, and therefore seeing a token only once, or simply having the knowledge that particular forms are possible is enough to assist in analysing and determining classes of stems.

5.7 Conclusion

We had designed an experiment that applied the linguistic criteria based on Evans and Osada (2005) in determining when certain word classes can be conflated. The results have shown that there are certainly distinguishable properties between nouns and verbs in Indonesian, even when we restrict ourselves by only examining features at the morphological level. The experiments used solely morphological features, showing promise that the labelling of word classes may be achieved without the use of collocational or syntactic features. These morphological pattern experiments look

promising in determining the class of unknown words or out of vocabulary items, as a means of extending the lexicons. For future experiments we would like to mix syntactic features with the morphological features used in this study. We would also like to extend the study to see how different the morphological patterns are from one source of text to another.

On a broader, more general note, this study has shown how issues in linguistics can be tackled using methods developed and used in the field of computational linguistics.

Chapter 6

Discovering Lexical Types

It has been shown that the verbal suffix *-kan* in Indonesian at times applicativises, licensing a benefactive object, at times it results in a causative construction, at times does not affect the valency of the affixed verb but shifts the prominence of the complements, and at other times seemingly reduces valency (Kroeger 2007; Son and Cole 2008). In Chapter 4, we described lexical types that we discovered as means to restrict the overapplication of *-kan* in the Indonesian deep grammar. These types are useful for us because annotating a lexical entry with this label maps out the possible changes to a verb's predicate argument structure (its alternations) when affixed with *-kan*. In this chapter, we conduct an experiment to see if we can automatically find stems that group into these types. In terms of defining our task, we aim to cluster groups of stems that share the same alternations when affixed with *-kan* in an effort to expedite lexicon development. We adhere to the notion that semantic similarity aligns with syntactic similarity as per Levin (1989). It has been shown that syntactic similarity has been used effectively to group semantically-similar verbs (Schulte im Walde 2009; Sun *et al.* 2010), but it has been shown to be less effective using semantic similarity to determine syntactic features (Baldwin 2005). Nonetheless, we employ a method that uses non-parametric Bayesian models as a means to model distributional semantics in order to test if we can discover Indonesian syntactic features. In our clustering task, we test whether the types we define in Section 4.4 are appropriate for the semantically driven of automatic discovery we employ.

Having an automatically-generated list of words to add to the lexicon with relevant suggestions for syntactic annotation would greatly assist lexicographers in the development of the grammar. Having this relevant information for items lacking syntactic information to include in the lexicon would greatly increase the speed and add to the ease of lexicon development. Our method of defining stem types focuss more on changes to argument structure in relation to *kan*. These systematic changes to the argument structure that are triggered by *-kan*, we refer to as *-kan* alternations. We explore a method of inferring this kind of syntactic information for Indonesian verbs. We apply a method that models distributional similarity for a lexeme using a

topic model under the assumption that lexemes that are semantically similar also display syntactic similarility. We also approximate structural information by including function words (stopwords) in our model.

In this chapter, we aim to apply methods akin to deep lexical acquisition (Baldwin 2005; Baldwin 2007). However, as described in Section 2.6, much of the work in this area requires the use of NLP tools and resources that are not available for Indonesian. For example, research conducted in classifying Levin classes employ subcategorisation features within the systems they build (Schulte im Walde 2006; Sun *et al.* 2010). In an experiment the authors consider fairly light-weight, Joanis *et al.* (2008) employ a part-of-speech tagger and a chunker. However, in Indonesian there are no part-of-speech taggers or chunkers that have enough training data to be used in large scale applications. For this reason, we do not and cannot rely on such NLP tools for our task, and therefore experiment with testing Levin's (1993) hypothesis by employing a semantic approach in finding subclasses of verbs that behave in a similar way syntactically.

In Section 6.2 we describe our gold standard data for evaluation, which involves the mapping of the changes to the predicate-argument structure of verbs when *-kan* is attached , and in doing so we briefly restate our criteria in creating our gold standard data, as detailed Section 4.4. Section 6.3 gives the details of our methodology, and our interpretation of distributional similarity expressed in soft clusters derived using the hierarchical Dirichlet process (HDP) and in Section 6.4.2 we apply our method to unlabeled data in order to extend the lexicon we initially endeavoured to create.

The experimental results are presented in Section 6.4, and in Section 6.5, we conduct a reanalysis to ascertain if the method we employ is suitable for the task of discovering syntactic information, or whether the discovery of synonyms modelled on Levin classes is more suited to this method. We finally conclude with how we may extend this preliminary investigation.

Our contributions in this work are: (1) the demonstration that hierarchical Dirichlet processes, in the family of Bayesian generative topic models, are a highly effective way of modelling word similarity, outperforming simpler strategies; (2) the application of the syntax–semantics hypothesis of Levin to an under-resourced language based on distributional similarity analysis; (3) conflating semantic classes into superordinate types may be useful for annotating the lexicon, but when performing clustering tasks that employ distributional semantics, having a more semantic oriented classification, such as Levin classes, are better suited for such methods, even when approximations are made to account for syntactic information; and (4) the demonstration that clustering based derived semantic properties has the potential to be good predictor of deep syntactic lexical properties, and this work is an initial step in semi-automatically constructing a deep lexical resource for an under-resourced language

6.1 Motivation

This research aims to help expedite lexicon development for precision grammars. At the inception of such projects there are not many resources readily available, and much of the research in verb classification relies heavily on curated linguistic information and natural language processing tools, such as treebanks, part-of-speech taggers, (dependency) parsers, and semantic knowledge bases, many of which do not exist for so-called low-density languages. Even experiments deemed to be 'light-weight' tend to minimally require a part-of-speech tagger and chunker (Joanis *et al.* 2008).

Deep language resources are often constructed on the basis of pre-existing resources, such as PropBank (Palmer *et al.* 2005) which was constructed over the Penn Treebank (Marcus *et al.* 1993), or the SALSA Corpus (Burchardt *et al.* 2006) which was constructed over the TIGER Treebank (Brants *et al.* 2002). For under-resourced languages, such resources tend not to exist, meaning one is often overwhelmed with obstacles at the outset. Sadly, such languages tend to fall by the wayside in this rich-get-richer reality of resource creation for NLP. This study investigates a means of deriving syntactic information with little more than a text collection and the hypothesis that syntactic distinctions often correlate with semantic distinctions.

The practical goal, which initiated this research, was to assist in developing an Indonesian lexicon for a deep grammar. And although these linguistically-motivated grammars are invaluable resources for the NLP community, the biggest drawback is the time required for the manual creation and curation of the lexicon. Our work aims to expedite this process by assigning syntactic information to stems that make up the verbal elements, on the basis of the predicting of alternation clusters based on distributional similarity.

Levin (1989) shows that there is a correlation between the semantics of a verb and its syntactic profile, however in this study we aim to see if we can induce groups of stems that behave the same way syntactically even if not all the stems grouped together are synonyms of each other or semantically related. The way we accommodate for this in modelling our experiments is by maintaining the stopwords when we gather contextual information. Stopwords are a good indication of syntactic structure. In preserving the stopwords as a proxy for syntactic modelling, and having surrounding content words for semantic context, we aim to see if this combination will suffice in grouping the manually constructed *kan* types we had defined. The groups are a little broader in membership than Levin classes – they are supersets of Levin classes, meaning that within the types we define will be subsets of semantically related stems. One advantage our stem types (or hyper-categories) have over the discrete Levin classes for our task of labelling items in a lexicon, is that it would result in fewer types to annotate in the lexicon. This would have the same usefulness being able to label a lexical item *ambitransitive* – we know it can be *intransitive*, and *transitive*, but not *ditransitive*. This label has no reference to the semantics of the stem,

and has immediate applications in parsing, showing this information focusing mainly on arity assists with syntactic parsing. For example, Briscoe and Carroll (1997) and Carroll and Fang (2004) enhance the performance of a deep grammar for English by extracting subcategorisation frames of lexical items, from corpora, with no reference to their lexical semantics. However, the distinctions we make are more informative than labels such as *ambitransitive*.

The aim of this study is to test if we can learn these syntactic alternation that are imposed by the affixing of *-kan*. We test the viability of inferring syntactic information from semantics, while accommodating for syntactic structure. We test our Levin-style assumption to see if the classes we induce have similar syntactic characteristics when compared to our manually created gold standard types using syntactic and semantic criteria. We aim to see if we can achieve what Levin showed – finding that semantically similar words behave in syntactically similar ways – and whether the method we propose is suited to this task.

6.1.1 Under-resourced Languages

Although Indonesian is spoken by some 23 million speakers as their mother tongue, and in excess of 165 million speakers throughout Indonesia (where it is the national language) and around the world (Gordon 2005), it still is an under-resourced language when it comes to NLP, as we discuss in Chapter 2. There is no robust part-of-speech tagger available to the community. There have been some successes in creating prototypes (Pisceldo *et al.* 2009; Wicaksono and Purwarianti 2010), but none have yet been integrated in broader NLP applications or reused in other studies, possibly because there is no standard part-of-speech tagset, or even a de facto tagset as there is for English with the Penn treebank (Marcus *et al.* 1993). There have been efforts in creating morphological analysers (Uliniansyah *et al.* 2002; Asian *et al.* 2005; Pisceldo *et al.* 2008; Mistica *et al.* 2009; Larasati *et al.* 2011), but because Indonesian is a relatively newly studied language in the NLP world, it is difficult to gauge the impact of these tools.

6.2 Gold Standard Data

We aim to group our stems according to the collection of their possible syntactic alternations with respect to *-kan*. For our gold standard data, we use the manually discovered alternation types in Section 4.4, where we also outline the method by which we arrive at these classes. As a reminder of the criteria used in the forming of these classes, we have collated the 8 verb types and presented them in Table 6.2 (see Appendix C.2 for a description adjective classes, and Appendix C.3 for nouns). We also present the 26 syntactico-semantic types with 8 verb, 13 adjective, and 5 noun classes in Table 6.1.

1.	V_1	*acuh* "to heed", *terjemah* "translate", *mandi* "bathe"
2.	V_2	*bawa* "carry", *beri* "give"
3.	V_3	*dengar* "hear", *kenang* "think of"
4.	V_4	*hidup* "be alive", *jatuh* "fall", *mati* "die", *tewas* "perish", *pusing* "to concern oneself", *minggir* "put aside", *masuk* "enter", *hadir* "be present", *lulus* "go through"
5.	V_5	*serah* "surrender", *singkir* "get out of way", *susup* "duck down"
6.	V_6	*bangun* "form/take shape", *pecah* "be broken"
7.	V_7	*paksa* "force", and also *buat* "make/do"
8.	V_8	*timpa* "hit", *baca* "read"
9.	N_1	*administrasi* "administration", *instalasi* "installation", *legalisasi* "legalisation", *nasionalisasi* "nationalisation", *ikat* "cord", *pukul* "blow/strike", *sewa* "hire"
10.	N_2	*ajar* "lesson"
11.	N_3	*gambar* "picture"
12.	N_4	*aplikasi* "application", *ekspresi* "expression", *kerja* "activity/work"
13.	N_5	*belanja* "expenses", *gelembung* "bubble", *buku* "book", *publikasi* "publication", *radiasi* "radiation", *kumandang* "echo"
14.	N_6	*darat* "land", *didih* "boiling"
15.	N_7	*hipotesis* "hypothesis", *titah* "command", *mimpi* "dream", *pikir* "idea", *tanya* "question"
16.	N_8	*asumsi* "assumption", *umpama* "example", *wakil* "proxy", *lokasi* "location"
17.	N_9	*paten* "patent" *tempat* "place" *tumpu* "foothold" *letak* "position" *penjara* "jail" *rumah* "ho
18.	N_{10}	*injeksi* "injection", *kait* "hook", *analogi* "analogy" *maklumat* "declaration"
19.	N_{11}	*sesal* "regret", *susu* "milk"
20.	N_{12}	*janji* "promise", *cerita* "news"
21.	N_{13}	*sarang* "web", *percik* "stain", *mula* "start", *kerja* "work"
22.	A_1	*abadi* "eternal", *asing* "separated", *cemar* "dirty", *cerdas* "intelligent", *goyah* "unstable", *haram* "prohibited", *murni* "pure", *mutakhir* "recent/up-to-date", *padu* "compact/solid", *populer* "popular", *salah* "wrong", *subur* "fruitful", *terang* "clear"
23.	A_2	*biasa* "ordinary/common", *unggul* "excellent/ahead", *berani* "audacious"
24.	A_3	*cengang* "amazed", *takjub* "surprised"
25.	A_4	*kecewa* "disappointed", *leceh* "worthless", *remeh* "unimportant", *teguh* "strong", *jengkel* "annoyed"
26.	A_5	*lunak* "soft", *lanjut* "protracted"

Table 6.1: All types

MORPHOLOGY	VERB FRAME	DECOMPOSITION
Type 1: *acuh* "to heed", *terjemah* "translate", *mandi* "bathe"		
MEN+V_1	–	–
MEN+V_1+KAN	<NP_a, NP_b >	DO_{to}(NP_a, [V_1 TO_{do}(NP_b)])
Type 2: *bawa* "carry", *beri* "give"		
MEN+V_2	<NP_a, NP_b > {PP_c}	DO_{to}(NP_a, [V_2 TO_{do}(NP_b) {[$_{path}$ P_c(NP_c)]}])
MEN+V_2+KAN	<NP_a, NP_b > {PP_c}	DO_{to}(NP_a, [V_2 TO_{do}(NP_b) {[$_{path}$ P_c(NP_c)]}])
	<NP_a, NP_b, NP_c >	DO_{for2}([NP_a], [V_2 TO_{do1}(NP_c) FOR_{do2}(NP_b)])
Type 3: *dengar* "hear", *kenang* "think of"		
MEN+V_3	<NP_a, NP_b >	$HAPPEN_{to}$(NP_b, [V_3 TO_{happen}[NP_a])])
MEN+V_3+KAN	<NP_a, NP_b >	DO_{to}([NP_a], [V_3 TO_{do}([NP_b])])
Type 4: *hidup* "be alive", *jatuh* "fall", *mati* "die", *tewas* "perish", *pusing* "to concern oneself", *minggir* "put aside", *masuk* "enter", *hadir* "be present", *lulus* "go through"		
MEN+V_4		
MEN+V_4+KAN	<NP_a, NP_b >	CAUSE([NP_a], [$HAPPEN_{to}$([V_4 TO_{happen}([NP_b])])])
Type 5: *serah* "surrender", *singkir* "get out of way", *susup* "duck down"		
ME+V_5+N	<NP_a >	DO([NP_a], [V_5])
MEN+V_5+KAN	<NP_a, NP_b >	CAUSE([NP_a], [$HAPPEN_{to}$([V_5 TO_{happen}([NP_b])])])
Type 6: *bangun* "form/take shape", *pecah* "be broken"		
MEN+V_6	<NP_a, NP_b >	CAUSE(NP_a, [$HAPPEN_{to}$([V_6 TO_{happen}(NP_b)])])
MEN+V_6+KAN	<NP_a, NP_b >	CAUSE(NP_a, [$HAPPEN_{to}$([V_6 TO_{happen}(NP_b)])])
Type 7: *force* "paksa", and also *buat* "make/do"		
MEN+V_7	<NP_a, NP_b > {VP_c}	DO_{to}(NP_a, [V_7 TO_{do}([NP_b]) {[DO([NP_b], [VP_c])]}])
MEN+V_7+KAN	<NP_a, NP_b, VP_c >	DO_{to}(NP_a, [V_7 TO_{do}([NP_b]) [DO(NP_b, [VP_c])]])
	<NP_a, VP_b >	DO_{to}(NP_a, [V_7 [DO_{to}([VP_b])]])
Type 8: *timpa* "hit", *baca* "read"		
MEN+V_8	<NP_a, NP_b >	DO_{to}(NP_a, [V_8 TO_{do}([NP])])
MEN+V_8+KAN	<NP_a, NP_b > {PP_c}	DO_{to}(NP_a, [V_8 TO_{do}([NP]) {[$_{path}$$P_c$([$NP_c$])] })

Table 6.2: Verb Types

These classes are not exactly the same as Levin classes because stems in a given class are not necessarily synonyms of each other. However, these types often form subgroups within a class that are synonymous and can be equated to Levin classes. For example, we can identify the Levin subgroup labelled as disappearance-48.2 in VerbNet within the verb type V_4 in Table 6.2, with the members *tewas* "perish" and *mati* "die". Also, *singkir* "get out of way" and *susup* "duck down" form their own subgroup, which can be labelled as the Levin class avoid-52.

6.3 Method

Our desired outcome is to cluster like stems on the basis of the distribution of their surrounding unigrams. We define our features in terms of the context of occurrence of our target lexeme, and employ hierachical agglomerative clustering over these features in two ways: (1) directly over the raw word frequencies; and (2) over extracted semantic features learned via the contexts of occurrence that we induce using a topic modelling approach.

These context unigrams represent the arguments for each instance of our target verb, and we employ HDP to discover the kinds of arguments that this verb normally takes.

We use Wikipedia as our text collection.[1] We removed mark-up with Wikiprep.[2] and tokenised with the English-trained models of OpenNLP[3] The total word count of the text collection was approximately 26 million words. In the next section we outline the features we use in our experiments, in addition to outlining our clustering method.

6.3.1 Feature Engineering

The features we employ are dependent on a number of filters or variables we define. These filters determine how we collect our unigram features from our text collection. We use three major filters in our this clustering task:

1. Morphological Filter; **morph** $\in$ 'k', 'mk', 'smk'
2. Window Size; **win** $\in$ 1 to 5
3. Context Filter; **context** $\in$ '+' (forward), '–' (backward)

[1] `http://dumps.wikimedia.org/idwiki/`
[2] `http://www.cs.technion.ac.il/~gabr/resources/code/wikiprep`
[3] `http://opennlp.apache.org/`

Morphological filters (morph): This filter determines if we collect contextual features for different morphological forms of the target lexeme, where: *'s'* stands for *stem*, i.e. the unaffixed lexeme; *'m'* stands for the AV variant of the lexeme, based on pattern (1) from Section 4.4; and *'k'* stands for the *-kan* suffixed form of the AV variant of the lexeme, based on pattern (2) from Section 4.4. An example of the *'s'*, *'m'* and *'k'* variants of *beli* "buy" are *beli*, *membeli*, and *membelikan*, respectively. These morphological filters determine whether the unigram features we collect for a lexeme is based on instance of s/m/k forms found in the text. We experiment with the context features based on these morphological variants in isolation and also in combination. For example, 'mk' would capture context features for the *membeli* and *membelikan* variants of the stem *beli* "buy". Note that we always include the 'k' features, as we are interested in changes induced by *-kan*, and we introduce 's' and 'm' features to determine whether they assist in classification, or simply add noise.

Window Size (win): This stipulates the context window size, relative to individual occurrences of the target lexeme, and can take a value from 1 to 5. A *win* of 3 would only gather unigrams up to (and including) three words away from the target word. We maintain stopwords in this experiment as a proxy to modelling syntax, because closed class categories, such as prepositions, are a good indication of syntactic structure.

Context Filters (context): We look at the backward ('–') or forward ('+') context unigrams. For *morph* = 'm' and 'k', preceding words will tend to capture the subject of the target lexeme, and following words will tend to capture the object of the target lexeme.

6.3.2 Clustering Stems

We employ hierarchical agglomerative clustering (HAC) in two ways: (1) over the raw frequencies of words based on a given feature representation defined in Section 6.3.1; and (2) over the output of the distributional semantic modelling (HDP) discussed in Section 6.3.3. The output of this step produces topic models. In other words, we perform HAC over raw unigram frequencies and induced topic models from these raw frequencies to ascertain the usefulness of the HDP step.

HAC[4] is a bottom-up clustering algorithm summarised by Jain *et al.* (1999:p277) in these three steps:

[4]We use HAC because it is commonly used in verb clustering tasks such as Schulte im Walde (2006); Sun and Korhonen (2011), and also because it is assumed that such verb classes (for example the representation in WordNet) exhibit a hierarchical structure. However, in these experiments we do not evaluate them as such.

1. Compute the proximity matrix containing the distance between each pair of patterns. Treat each pattern as a cluster.

2. Find the most similar pair of clusters using the proximity matrix. Merge these two clusters into one cluster. Update the proximity matrix to reflect this merge operation.

3. If all patterns are in one cluster stop. Otherwise, goto step 2.

To compute the distance between a pair of patterns, we use Squared Euclidean, and for the linkage criterion for merging clusters we use weighted linkage clustering (WPGMA).

We compare the output of HAC with the gold standard classes, and so enforce flat clusters from the hierarchical clusters by applying a distance threshold t.[5] This threshold determines whether an instance should be grouped within a cluster or not.

6.3.3 Modelling Distributional Similarity

Distributional semantic models are commonly employed in the induction and disambiguation of word senses (McCarthy *et al.* 2003; Lapata and Brew 2004; Brody and Lapata 2009; Lau *et al.* 2012), and to a lesser extent, in learning syntactic classes and diathesis alternation behaviour (Parisien and Stevenson 2011; Bonial *et al.* 2011). We infer lexical similarity and soft word clusters using topic modelling, based on a hierarchical Dirichlet process (HDP: Teh *et al.* (2006)), a non-parametric extension of latent Dirichlet allocation (LDA: Blei *et al.* (2003)).

The latent topics induced by these models take the form of a multinomial distribution over words in the document collection terms. They capture a coherent semantic aspect of the document collection, which is our modelling of distributional semantics for the purpose of our task. For this task, we define the 'document' as the selected context around our target *kan*-affixed verb. We define a small window around our target word as a means of approximating the kinds of arguments the target word allows.

6.3.4 Evaluation

We develop two baseline systems to compare our results against: (1) a majority class method that clusters all stems into one group; and (2) random class assignment based on a uniform class distribution over as many classes as there are in the

[5]We determine this threshold by conducting preliminary experiments. The set-up of these experiments is the same as those described in Section 6.4, but we use a subset of this data, which as approximately 10 million words. For these experiments we do not perform any cross-validation, but simply run the experiments allowing all possible values for all the variables in the experimental set-up. We simply choose the top 25 best performing systems and allow simple voting to determine t.

gold standard. The random scores reported are based on the median of 11 random assignments.

We use pairwise precision (pP), recall (pR), and F-score (pF_1) to evaluate our generated clusters, relative to the gold-standard word classes.

Using the method outlined in Manandhar *et al.* (2010), we convert the clustering task into a classification task. If (i) represents the number of items for the cluster C_i, then we produce $\binom{i}{2}$ number of pairs for that cluster.

This description is formulated by Menestrina *et al.* (2010) as follows and is described in more detail in Section 3.3.6.

$$pP(I,G) = \frac{|Pairs(I) \cap Pairs(G)|}{|Pairs(I)|}$$

$$pR(I,G) = \frac{|Pairs(I) \cap Pairs(G)|}{|Pairs(G)|}$$

$$pF_1(I,G) = \frac{2 \times pP(I,G) \times pR(I,G)}{pP(I,G) + pR(I,G)}$$

This evaluation metric is commonly used in unsupervised tasks (Manning *et al.* 2008), such as for word sense induction and disambiguation (Manandhar *et al.* 2010) (see Section 3.3.6 for more details). It has also been used in evaluating discovered Levin-style classes for German by Schulte im Walde (2006), an experiment comparable to our task.

6.4 Experiments

For these experiments, we model the distributional similarity step using a hierarchical Dirichlet process (HDP). These produce topic probabilities from the 735 stems identified above, over which we perform hierarchical agglomerative clustering (HAC) to induce our syntactico-semantic classes.

We perform 2 kinds of experiments: (1) ON-ALL; and (2) ON-LIKE. The difference between ON-ALL and ON-LIKE is that in the latter, we cluster (using HAC) over topic models extracted from stems that belong to the same part-of-speech. For example, ADJECTIVE ON-ALL results report on clustering (HAC) of adjective stems over the topics discovered (using HDP) using only features from adjective stems. On the other hand, ADJECTIVE ON-LIKE reports on clustering of adjective stems over topics discovered using features from stems from all parts-of-speech. That is the topics are extracted taking into account all 735 stems for the ON-ALL experiments, whereas ON-LIKE experiments will have had subset of of 735 from which the topics are generated. Therefore for ADJECTIVE ON-LIKE topics are generated from all the adjective stems

appearing in the list of 735 stems we had identified.

To ascertain the usefulness of the HDP step, we compare the classes induced using the above method with a method that does not employ the extracted topics. This method has been employed with systems developed by Schulte im Walde (2006) for German and Jurgens and Stevens (2010) for English word sense induction. In a similar way, we simply perform HAC over the unigram features we define in Section 6.3.1. These experiments we label NoHDP. The ON-ALL results for the NoHDP experiments (i.e. with no topic modeller step) are trained on all word classes and evaluated only on one word class, i.e. the NoHDP adjective ON-ALL take into consideration all word classes when clustering and evaluate only on adjectives. The results we label HDP have an added step of having the topic modelling performed over the raw unigrams, the results over which we perform HAC.

6.4.1 Determining Features

Our aim in this section is to define the filters we use in determining our features when we cluster (HAC) the 735 stems. Of these 735 stems, there are 100 labelled stems we had annotated as our gold standard, and the rest are unlabelled stems.[6] As we had seen in Section 6.3.1, we had 3 kinds of filters in extracting our features: morphological filter, window size and context (backward or forward).

We first perform clustering over our gold standard data to determine the optimal settings for our 3 filters. The backward and forward context filters represent the subject (backward) and object (forward), and it would be interesting to discover which of these are more predictive in determining our syntactico-semantic classes. We had hypothesised that for nouns we would discover that the backward contexts would be most useful in determining our classes, for verbs '+', and for adjective a combination of both '-' and '+'. The idea behind the window size was to see if function words rather than content words surrounding the lexeme in question were more important in this task; the larger the word window the more surrounding content words would be captured as features, and the smaller the word window the more likely only function words, such as prepositions that signal structure would be allowed as features.

Our predictions for these experiments in determining the optimised values for our filters, as shown in Table 6.3. The basis of the predictions for the verbs are taken largely from Jackendoff's (1996) theory that the expressing of a bounded object affects the inherent semantics of the verb and therefore its syntactic structure. [7]

[6] As noted in Section 2.6.1, the advantage of using unsupervised methods is the ability to still learn from unlabelled data given the expense of annotation.

[7] That is the object *the tune* renders the singing event telic in (b), while in (a) the verb is atelic.

a. Bill sang (*in five minutes).

b. Bill sang the tune (in five minutes).

This aspectual property of the Indonesian verb can be expressed with the presence of *-kan* as noted

	ON-ALL				ON-LIKE		
	A	N	V	ANV	A	N	V
HANDBUILT	k3-+	k3-	k3+	k3-+	mk3-+	mk3-	mk3+

Table 6.3: Handbuilt predictions.

We based our hypothesis of the noun handbuilt filter settings on the fact that many nominals when affixed with AV produced intranstive verbs,[8] and because the majority of the semantic changes we had observed in Section 4.4 (with the affixing of *-kan* to *meN* +stem) were not overwhelmingly causative, we had hypothesised that the backward context would be more informative. For the adjective stems, we had observed many resulting intransitive and transitive verbs from our classification of types in Section 4.4, and decided to allow for both the forward and backward contexts in the HANDBUILT filter settings.

We report on experiments on 10-, 5-, and 2-fold cross-validation in Table 6.4, for methods that include the topic modelling step (HDP) and those that did not (NO-HDP). We also employ a *bagging approach* (sampling with replacement) to ascertain the best parameters to apply to our 735 lexemes in terms of the unigram features we define in Section 6.3.1. Through this bootstrap aggregating procedure, we discover the optimal filters values and settings shown in Table 6.4. On the left-hand side of the table are the ON-ALL experiments, with topic models extracted from all parts of speech, but evaluated on the subset of 100 stems in the gold standard data, where ANV is the combination of all parts-of-speech.

The result for the NoHDP optimal feature discovery was rather inconsistent, however the findings for the HDP feature settings show that backward (–) context is the most informative for the ON-LIKE experiments. However the ON-ALL results show

by Son and Cole (2008), who claim that this affix renders a *kan*-affixed as signifying a resulting event. In Indonesian the suffix *-kan* captures the difference between (a) and (b), as shown in Examples (c) and (d).

c. Bill nyanyi.
B. sing
"Bill sang."

d. Bill menynanyikan lagu itu
B AV+sing-KAN song that
"Bill sang the song."

[8]Some examples are:

darat	"land"	*mendarat*	"to land"	A plane landed.
batu	"stone"	*membatu*	"to freeze"	He froze (like a stone).
tingkt	"level"	*meningkat*	"to rise"	The temperature rose.

	ON-ALL				ON-LIKE		
	A	N	V	ANV	A	N	V
HDP-10FOLD	mk3-+	mk3-	k1-	mk1-+	k1-	k1-	mk1-
HDP-5FOLD	mk2-+	k1-	mk3-	mk5+	k1-	k1-	mk1-
HDP-2FOLD	mk2-+	k1-	mk3-	mk5-	k1-	k1-	mk1-
HDP-BOOTSTRAP	mk2-+	k1-	mk3+	smk3+	k1-	k1-	mk2-
NODHP-10FOLD	mk5+	mk5-	mk2+	mk2+	mk5+	mk5-	mk2+
NOHDP-FOLD	mk2+	k5-	k5+	k5+	mk2+	k5-	k5+
NOHDP-2FOLD	k1-+	k5+	mk5-	mk2+	k1-+	k5+	mk5-
NOHDP-BOOTSTRAP	k3-	mk5+	mk5-	k5-+	mk2+	mk5+	mk4-

Table 6.4: Discovered filter settings for **morph**, **win**, and **context** for HDP and NoHDP

that the backward context is most informative for nouns, and verbs (except for the bootstrapping experiments), while having both surrounding contexts for adjectives were optimal. Except for ANV, the best window sizes were fairly small, which was a good indication that our resulting topics and their defining list of words (with the highest probability distribution) for these experiments would have a high proportion of structural indicators. Furthermore, as expected, we found a high number of function topics (topics with primarily function words), but upon inspecting these resulting topics, we found the same function words would occur in many of the function topics. This was a good indication that our modelling of structural indicators was not robust. The non-function topics (topics that had a list of items that seemed to be semantically related), were indeed semantically cohesive, as we expect from this the topic model. Nonetheless, we apply these optimised context settings to the 735 data, shown in Section 6.4.2.

6.4.2 Application of Discovered Context Features

Based on the best features determined by our filters derived from the 100 stems, we perform clustering over the 735 stems we identified earlier that can have *kan*-affixation. We evaluate only on the 100 stems we had created in our gold standard.

Overall, the HDP systems exceeded the performance of the NoHDP systems. If we compare the ON-ALL NoHDP systems with the HDP in Table 6.5, we see that the paired F-score for the NoHDP model performs considerably below that for the HDP model, showing that intermediate semantic modelling using HDP enhances accuracy.

Overall, the results of the HDP systems in Table 6.5 far exceeded the results of NoHDP, even though they did not convincingly outperform the majority class baseline (and for the adjective, the results were in fact below the Majority Class

	HDP							NoHDP						
	ON-ALL				ON-LIKE			ON-ALL				ON-LIKE		
	A	N	V	ANV	A	N	V	A	N	V	ANV	A	N	V
10-FOLD	.352	.126	.230	.077	.303	**.166**	.252	.015	.031	.017	.039	.165	.088	.211
5-FOLD	.458	**.162**	.237	**.098**	.303	**.166**	.252	.023	.024	.031	.025	.100	.035	.167
2-FOLD	.458	**.162**	.237	**.112**	.303	**.166**	.252	.022	.012	.027	.039	.124	.139	.156
BOOTSTRAP	.458	**.162**	**.281**	**.097**	.303	**.166**	.261	.032	.033	.026	.042	.100	.106	.152
HANDBUILT	.376	.141	.087	**.099**	.035	.057	.038	.035	.035	.012	.039	.017	.048	.156
Maj. Class	**.471**	.146	.271	.094	**.471**	.146	**.271**	**.471**	**.146**	**.271**	**.094**	**.471**	**.146**	**.271**
Random	.253	.077	.140	.044	.253	.077	.140	.253	.077	.140	.044	.253	.077	.140

Table 6.5: Results: Pairwise F-score

Baseline). Even though we see that the HDP results did convincingly outperform NoHDP, with F-scores only marginally exceeding the Majority Class Baseline, we cannot be certain that the learner is producing reliable classes. Also we had expected verbs to outperform all other word classes in these experiments because as was shown in Chapter 5, the morphological context ME+STEM+KAN suggests that the stem is likely to be a verb. However, we observe that nouns consistently bettered the results of verbs and adjectives in comparison with the baselines. One possible cause may be the small number of verbs in the gold standard data, especially once we break down the evaluation into word classes. The low number of stems in our gold class may also partly explain the lacklustre results.

There were approximately half as many verb stems as there were nouns, with only 25 in the gold set. Before embarking on more costly annotation, we wanted a way to verify the utility of the methodology. Assuming that the classes we induce with the method are stems that behave in the same way when affixed with PAS-changing morphology, then we should be able to reproduce that same stems groups for other PAS-changing morphology, such as the locative suffix *-i*, which behaves in a parallel way as *-kan*, as we see in Section 6.5.

6.5 Validating the Methodology

Given the lacklustre performance of our system, we verify the utility of the methodology used in this study by performing what we call the *i-X-kan* experiments in Section 6.5.1. If indeed the methodology described in Section 6.3 is useful in grouping *-kan* alternation classes, then this method should also be useful for other PAS-changing morphology, such as the locative suffix *-i* (Arka *et al.* 2009). The *-i* suffix exhibits a similar behaviour to *kan*, but instead of introducing a benenfactive

object (when the number of direct arguments are increased), this suffix introduces a locative object, as shown in Example (6.1), or shifts the locative to the direct object position, as shown in Example (6.2) (for a more thorough account of the -i suffix, see Arka *et al.* (2009)).

(6.1) [from (Arka *et al.* 2009)]

a. *Ayah mengirim uang kepada dia*
father AV+send money to him
"Father sent money to him/her."

b. *Ayah mengirimi dia uang*
father AV+send him money
"Father sent him/her money."

(6.2) [from (Arka *et al.* 2009)]

a. *Ia melepar batu ke saya*
(S)he AV+throw stone to me
"She/he threw stones to me."

b. *Ia melepari saya dengan batu*
(S)he AV+throw stone to me
"She/he threw stones to me."

If the clusters of stems we induce with this method have linguistic significance, then it should generalise beyond only predicting the behaviour of *kan*-affixation. Although we find results confirming the validity of the method, we do further analysis to find that the overwhelming bias in the data to not form many clusters skews our evaluation of the data to be unrepresentatively high. Instead we return to a more fundamental question about whether out method is more suited to learning distinctions at the semantic granularity determined by Levin classes than a definition that tries to cluster solely morpho-syntactic behaviour, even with the inclusion of function terms as proxy for structural indicators. The Levin classes experiments presented in Section 6.5.2 employs the identical methods to our original experiments. Given that we had more success in the application of this method to our Levin-style classes than our *kan* classes suggests that this method is more suited to finding semantic distinctions rather than syntactic features.

6.5.1 The *i-X-kan* experiments

In these experiments, we gather 86 stems, which have attested instances with the suffix *-i*, and can also be suffixed with *-kan* (although not at the same time). These

	ON-ALL				ON-LIKE		
	A	N	V	ANV	A	N	V
80-HDP-100	.462	**.146**	**.455**	**.125**	.469	.140	**.397**
	mk3-	k1-	smk1+	smk2+	k3-+	k2+	smk3+
Majority Class	**.471**	**.146**	.271	.094	**.471**	**.146**	.271

Table 6.6: Discovered filter values for **morph**, **win**, and **context** bootstrap

are not an exact subset of the 100 gold stems we have above, with only a handful of stems from our 100-stem gold data that have attested usages of both the suffix *-i* and *-kan*.

We use the same experimental set up as described in Section 6.3, but instead of comparing the resulting clusters against gold standard data, we calculate agreement between the clusters induced from the *-i* data and the *-kan* data. To create our *-i* data, the *-i* context, morph, and win(dow) filter values are determined from experiments using Wikipeda, as described in Section 6.3.1 for *-kan*, but instead of collecting context features based on the morphological filter *smk*, we target *smi* morphs (for *-i* suffixes).[9] The *-kan* data is collected as usual, but only for the 86 stems.

Another difference in this experimental set up is that we only cluster over topics formed by the 86 stems and not all the 735 stems as we had done in the previous experiments. This is because we are not aiming to apply any of our findings to unknown stems, but we are only aiming to discover whether this method will induce the same kind of clusters for both sets of data.

One problem we face is determining what parameter settings to use for *morph*, *context* and *win(dow)*. To determine this, we use the unigram features from the 100 gold stems to induce our topics over which we induce clusters using HAC. There is not a complete overlap with the stems within the 100 gold data and the 86 *i-X-kan* data, but they at least represent a similar number of stems from which we can learn topics, over which we cluster the stems.

The learned parameters from our bootstrapping experiment using the 100 gold stems are shown in Table 6.6, with the pF-score from the bootstrapping experiment reported on the top line, and the discovered optimal settings on the second. For these experiments, we see that only V ON-ALL, ANV ON-ALL, and V ON-LIKE exceed the majority class baseline.

We use these learned parameters to induce clusters from our 86 stems both for the *-i* data and for the *-kan* data. From these induced clusters we calculate Cohen's

[9]The *smi* and *smk* describe the morphological affixes on the surface word of the lexical item of interest. The *s* is the bare stem, *m* is the stem with the *meN-* prefix, and the surface word of *k* items have both the prefix *meN-* and suffix *-kan*.

	ON-ALL				ON-LIKE		
	A	N	V	ANV	A	N	V
Kappa	.214	–	.843	.958	–	.554	.852

Table 6.7: Kappa values to test agreement between clusters in induced with *-i* and *-kan* data

Kappa agreement score, reported in Table 6.7. Kappa is calculated based on pairings within an induced cluster. For example if we had two systems A, and B, and had items w, x, and y that were clustered in the following way:

system	cluster 1	cluster 2
A	w, x	y
B	x, y	w

The pairings on which we calculate our Kappa score would look like the following:

Pairs	A	B
w+x	yes	no
w+y	no	no
x+y	no	yes

The '–' in the table indicates that either or both of the *-i* induced clusters or *-kan* induced clusters had found their way into only one cluster. Artstein and Poesio (2008) report that previously Kappa values above 0.6 were substantial.[10] However Artstein and Poesio's (2008) recent study found that manual annotation for computational linguistic tasks ensured reliable quality above a Kappa score of 0.8.

The scores in Table 6.7 suggest that the two systems have very high agreement in their induced clusters for V and ANV ON-ALL and for V ON-LIKE. However upon inspection of the clustered stems, and of the raw figures for V ON-LIKE in this contingency table, we see that the high number of disagreements artificially boosts the Kappa score, suggesting a higher agreement than there is. We had initially thought that this high score was due to Cohen's Kappa overly rewarding any agreement between annotators, or in this case our two systems, when the expected agreement is low, as in this case (with the possibility of random agreement Pr(e) ¡ 3%). This is indeed contributes, but the sytems readily form many clusters with few members, which means when we calculate Cohen's Kappa on pairwise agreement, there are many pairs that would not be found in both systems, even though there are far too few agreement on positive pairs found for the metric to be useful in this instance.

[10]However, they report this figure from a non-computational linguistic study.

		-kan	
		yes	no
-i	yes	2	27
	no	33	403

We conclude that this method of gauging similarity is not appropriate in comparing two automatic systems, but we cannot conclude from these results that the method itself is not useful for the task of inducing sytactico-sematic classes. One reason for the lacklustre results seen in Section 6.4.2 using this experimental method may be due to the gold standard data being too coarse, as we had previously suggested. In the following section we test this hypothesis by seeing whether we can induce Levin verb classes based on VerbNet.[11]

6.5.2 Reassessing Verb Types: Experimenting with Levin-classes

To test the hypothesis that our method yields less than optimal results because such a method is not suited to the task, and may be more amenable to finding semantic rather than syntactic features, we conduct a pilot study with a number of verb stems. We form our gold standard data using synonyms from VerbNet 3.2[12] as our guide in forming Levin classes for Indonesian.

From the verb stems in our data, we create Levin-style classes based on the translation of the senses in VerbNet. Because of the small number of Indonesian verbs in our gold standard list, we find translations for in our original list of verbs in this English resource, we expand the number of items we have for this task. These are shown in the *Additional Verbs* column in Table 6.8. The way we add extra verbs is also through synonyms that we could find in the list of 735 stems we identified in Section 4.4. However, even if intuitively two verbs (from their translations) should be in the same class, if they are not attested in VerbNet we do not use them. For example *kirim* "send" seemed as though it should be grouped with *beri* "give", but according to VerbNet *send* belongs to **send-11.1** and **instr_communication-37.4**, among other classes, while *give* belongs to **give-13.1** plus some senses that involve the verb give in multiword idiomatic expressions, such as *give_birth* and *give_in*.

We separate out our manually created verb types (V_1, V_2 etc.) from Table 6.2 into their Levin-style component parts, and only include stems from our original list if we were able to find the appropriate VerbNet class. Otherwise, we omit these from the experiment. There is no VerbNet in Indonesian, and so we construct the Indonesian Levin classes, for this experiment, on the translation of the appropriate sense of the verb for each VerbNet category. Also, given that VerbNet only pertains to verbs, we perform experiments only on the verbs.

[11] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

[12] http://verbs.colorado.edu/~mpalmer/projects/verbnet/downloads.html

Original Verb(s)	Additional Verbs	VerbNet Class
acuh "take heed"	–	–
terjemah "translate"	–	turn-26.6.1
mandi "wash"	–	dress-41.1.1
bawa "carry"	–	carry-11.4
beri "give"	*jaja* "hawk/sell", *pinjam* "lend"	give-13.1
dengar "hear"	–	–
kenang "think":	*kenal* "know" *ingat* "remember"	consider-29.9
hidup "be alive"[13]	–	–
jatuh "fall"	*panjat* "climb", *naik* "rise/climb", *roboh* "fall"	calibratable_cos-45.6
jatuh "fall/collapse"	*runtuh* "disintegrate"	other_cos-45.4
mati "die", *tewas* "perish"	–	disappearance-48.2
pusing "be concerned"	–	–
minggir "put aside"	*alih* "move to another place"	put_spatial-9.2
hadir "be present"	–	–
lulus "go through"	*maju* "advance/progress"	succeed-74
lulus "go through"	*tamat* "finish"	complete-55.2
masuk "enter"	–	–
serah "surrender"	–	–
susup "duck down", *singkir* "get out of way"	–	avoid-52
bangun "form"	–	–
pecah "be broken"	–	–
paksa "force"	–	force-59
buat "make/do"	–	–
timpa "hit"	*hantam* "hit/blow" *tabrak* "hit"	hit-18.1
baca "read"	*tulis* "write"	say-37.7
baca "read"	*hafal* "memorize"	learn-14

Table 6.8: Levin Classes

System	Majority Class	Random	ON-ALL	ON-VERBS
LEVIN-HDP			**.174**	**.367**
LEVIN-HANDBUILT	.114	.065	.086	**.136**
LEVIN-NOHDP			.057	.111
TYPES-HDP			**.281**	.261
TYPES-HANDBUILT	.271	.140	.087	.057
TYPES-NOHDP			.026	.152

Table 6.9: pF_1 score comparing benchmark system NOHDP with our HDP system for Levin Classes (LEVIN) and our coarser-grained TYPES

If a stem belongs to more than one VerbNet class, then we have added these as separate classes in our evaluation data. For example, *baca* "read" belongs to both **say-37.7** and **learn-14**, and we include both senses as separate classes. Each line in Table 6.8 represents one VerbNet classes, unless indented, and each line grouped together with a horizontal line represents the original gold classes (TYPES). As mentioned earlier, those verbs that have no VerbNet Class associated with them are omitted from the experiment.

We find the optimal parameter settings for *morph*, *context*, and *win(dow)* as usual, and find that for ON-ALL we have smk, '+' and 1, respectively, which means that for a lexeme, we collect context unigrams 1 word from our target lexeme in the forward context, and the morphological surface forms of our target lexeme, is all three forms described in Section 6.3.1. For ON-LIKE we discover the best settings are mk, '-+', and 3. These results applied to all verbs in the 735 stem data are reported in the HDP-LEVIN row in Table 6.9. The NOHDP-LEVIN results report on the parameters smk, '+', and 1 for both ON-ALL and ON-LIKE on the 735 stem data. Again, we find the pattern that including the HDP step vastly improves results over NoHDP, which in this case does not even surpass the Majority Class baseline.

The stems *mati* "die", and *tewas* "perish" from V_4 form ready-made Levin classes for **disappearance-48.2**. The verbs *timpa* "hit" and *baca* "read" from V_8 belong to the classes **hit-18.1** and **say-37.7**, respectively. Where we could, we added extra items to Indonesian Levin classes we formed.

In Table 6.9 we also repeat the original results reported in Section 6.4.2, for comparison. We see that our Levin experiments perform better than the original experiments using the *-kan* alternation classes we devised. The TYPES-HDP just passes the Majority Class baseline, or fails to exceed it.

A	*main* "play", *nyanyi* "sing", *gesek* "scrape"
B	*kirim* "send", *hantar* "place"
C	*terbang* "fly", *lempar* "throw"
D	*dapat* "get", *menang* "gain/win", *terima* "receive"

Table 6.10: Induced groups with no known categorised words

6.6 Discussion

The improved performance of the experiments using Levin classes in Section 6.5.2 show that inferring structrual information simply by including function words does not suffice. One possible way of improving our feature engineering to accommodate syntactic information is to include position information along with the unigram item as a double-barrelled feature type, such as Sun *et al.* (2010) and Lau *et al.* (2012). We aimed to achieve the learning of this positional information with our variable length window, but our method is not as precise as explicitly labelling position. The upside to this experiment is the finding that Levin classes are learnable using this method, and although they are not exactly equivalent to the types we have encoded, they can be mapped very easily, because there is a many-to-one mapping between Levin classes and IndoGram *-kan* types.

6.6.1 Analysing Induced Levin Classes

In this section we examine a small sample of the resulting stem groups from the Levin Class experiments. Table 6.10 shows membership of all stems found in four separate clusters. These 4 particular groups do not have any of the original 100 stems as members, unlike the groups formed in Table 6.11. In this table, the top half are groups that match our Levin Classes presented in Table 6.8, and the bottom half are examples of those that do not.

Group A from Table 6.10 has 3 verbs, *main* "play", *nyanyi* "sing", and *gesek* "scrape", which may initially seem not to form a semantically coherent group, however they are all associated with producing music. The verb *main* "play" is used to describe the playing of most musical instruments, and *gesek* "scrape/rub" is used for string instruments, such as violins, or cellos. Group B has members that describe movement from one place to another, as does Group D; and both members of Group C describe some projection into the air.

Groups F and G in Table 6.11 faithfully replicate the Levin Classes **avoid-52**, adn **learn-14** from Table 6.8. However, Groups H and I seemed to not form very coherent semantic groups.

F	*singkir* "get out of way", *susup* "duck down"	–
G	*baca* "read"	*hafal* "memorise"
H	*terjemah* "translate"	*tulis* "write", *muat* "insert/contain"
I	*paksa* "force"	*pinjam* "lend" *hapus* "wipe off/vanish/blot out"

Table 6.11: Induced groups with known categorised words

6.6.2 Word Class Analysis

From the results in Section 6.4.2, it was surprising that nouns performed better than the other parts-of-speech, given that predicting the semantics of denominalised verbs has been shown to be rather idiosyncratic. For example Jackendoff (2002:p.35) notes that the interpretation of the following denominalised verbs in English are conventionalised and not entirely predictable:

	"put N on"	butter, water, paint, roof
	"take N off"	dust (the shelves), scale (a fish), skin (a cat)
(6.3)	"put on N"	saddle (a horse), shelve (the books)
	"put in N"	pocket (the money), bottle (the wine)
	"fasten with N"	glue, staple, nail, tape

We had expected these experiments to perform best on verbs (V ON-ALL and V ON-LIKE) because as shown by Mistica *et al.* (2011), affixing *kan* with the prefix *meN* is a morphological pattern that a verb normally slots into. However, the idiosyncratic behaviour may have been due to there being more noun instances than verbs in the training data.

6.7 Conclusion

We have explored the question of whether distributional similarity models can be used to learn deep syntactic features for an under-resourced language, namely Indonesian. Our results demonstrate that hierarchical Dirichlet processes are a highly effective way of modelling word similarity, and outperform a simpler strategy of simply applying HAC over raw frequencies. We have also shown that learning classes geared toward the potential morpho-syntactic alternations of stems while conflating the the semantics of the stem are not amenable to this particular method. The pilot study that used true Levin classes for evaluation performed much better in comparison to the baselines than the experiments where we induced our manually-constructed types. We would need to model syntactic structure more effectively to gain better success in predicting types rather than Levin classes.

From this pilot study, we have applied the syntax–semantics hypothesis of Levin to an under-resourced language, namely Indonesian, based on a distributional similarity analysis. The results show that although the two stage method of initially inducing topics using HDP outperformed the HAC method, we still need to improve the precision of the system in order show that this method would be of assistance to a lexicographer in the semi-automatic construction of a deep lexicon.

Part IV

Concluding Remarks

Chapter 7

Conclusions

This thesis examined aspects of deviant morphology in Indonesian. In particular, we investigated Gil's (1994, 2001, 2005, 2010) claim that Indonesian dissolves the distinction between open classes categories. However, our study verified the need to maintain word classes in Indonesian from a linguistic perspective, and showed that word class induction can be successfully applied using only morphological features.

We detailed our contribution to a precision grammar with the encoding of the multi-faceted *-kan*, and mapped out the possible alternations of *-kan* with respect certain stems. These collection of possible syntactic alternations relevant to certain stems formed our defined *kan* types, designed to mitigate overgeneration in the lexicon. We embarked on a case study to determine whether these types could be learned using a semantic model, which in effect allows a semantic model to predict syntactic behaviour. As reported in Sections 2.5 and 2.6, the hypothesis that syntactic similarity aligns with semantic similarity has been widely investigated for a number of languages. However, as has been shown by Baldwin (2005), and in this study, using semantic similarity to discover syntactic features has not yet been overwhelmingly successful. However, we found that our method better suited discovering Levin classes, which we could potentially exploit in discovering syntactic features.

7.1 Future Work

In the word classes study we had undertaken, we had a specific linguistic question in mind, and actively sought out an answer. However, in the creation of linguistically motivated tools and resources, answers to linguistic questions present themselves without having being asked. For example, the manual creation of a Basque dependency treebank by Arantzabe-Urruzolak (2008), allowed the discovery of a canonical word order in Basque, even though it is described as a topic-focus language. From a grammar engineering perspective, by simply implementing linguistic analyses, we can refine the knowledge we have about the language under investigation as shown

by Bender (2009), as we had discovered in Chapter 4 upon constructing sublexical rules to account for the realisation of *voice* in Indonesian.

In terms of resource creation for IndoGram, we aim to continue to build up the verbal lexicon – at present there are only 150 verbs encoded in XLE lexicon, however there are approximately 2000 nouns. In addition to simply increasing the number of verbs in the lexicon, we also aim to better characterise the possible alternation classes to restrict overgeneration in the lexicon, and to apply templatic types to verbs. We also aim to extend the investigation to encompass *-i* suffixed verbs in our *-kan* investigation, so that we have a unified account of these predicate argument structure-changing affixes.

Alternation Classes

Our method which manually groups alternation classes, or types of stems that alternate in the same way with respect to *-kan* did not lend itself to a speedy development of inventory. Mainly because the semantic decomposition and the primitive we had defined in Example (4.4.4) were too inexact for the stems aimed to build up.

For future work we would like to improve the method by which we manually map out alternation types. Rather than discover the alternation types to begin with, we could take the same approach that was taken by Levin (1993), which results in classes with semantically-related members that alternate in the same way. Another alternative is to investigate semantic theories that may not employ primitives, but could be equally effective in our task, such as employing a semantic framework that has been applied to Austronesian languages namely the mode of lexical decomposition used in Role and Reference Grammar (van Valin and LaPolla 1997).

Improving DLA for Indonesian

The preliminary work in Chapter 6 suggests that more sophisticated techniques are required to model syntactic and semantic information that we were trying to learn in tandem. One line of investigation is rethinking the model we employ. In addition to devising more sophisticated features, we could employ more sophisticated algorithms that incorporate n-gram language models, such as the *Bigram Topic Model* (Wallach 2006), which extends Blei *et al.*'s (2003) latent Dirichlet allocation (LDA) model by incorporating it with Mackay and Peto's (1995) hierarchical Dirichlet language model. This hybrid model builds in the notion of word order, and also infers a separate singular topic for function words. Other hybrid systems that combine syntactic and semantic models develop composite models incorporating Hidden Markov models (HMM) and latent Dirichlet models (Griffiths *et al.* 2005), and HDP with HMM (Teh *et al.* 2006; Fox *et al.* 2009). These systems discover semantic topics in addition to syntactic functional phrases.

In addition, for future work, it would be interesting to exclude the stop words[1] in the HDP experiments, both for the experiments that were evaluated against the Levin-style classes and for experiments that used the types we defined around the suffix *-kan*. Doing this kind of experiment would show what kind of effect these stop words had in being able to emulate some sort of syntactice structure, but for the Levin-style experiments then it may reduce unwanted interference from these high frequency words that lack of semantic content.

Revisiting Issues Surrounding Resources

Li and Brew (2008) found that collocation features are useful in arriving at Levin classes, which means this reduces the need to preprocess the data in order to utilise syntactic and other linguistic information that is often used in Natural Language Processing. Even so, under-resourced languages do not have as much (raw) data to utilise effectively unlike the English study executed by Li and Brew (2008). It has also been shown for German (Schulte im Walde 2006) and French (Sun *et al.* 2010) (and even for English (Sun *et al.* 2008)) that syntactic features have proven to be useful in the determination of verb classes.

One way that we may be able to introduce syntactic features for future work is to employ methods involving parser transfer and transfer learning from English. These are learning methods where poorly resourced languages leverage the data and tools developed for well-resourced languages (McDonald *et al.* 2011s; Täckström *et al.* 2012; Naseem *et al.* 2012). Although English and Indonesian are not related languages. They do have superficial word order similarities for simple declarative active sentences.

7.2 Final Remarks

The definition of Computational Linguistics that we provided in Chapter 1 was that it was "trying to do what linguists do in a computational manner." What linguists do is build models that best describe language, however, as Corbett *et al.* (2002:95) state "computational linguistics can provide a means of *validation*, of checking whether a particular theory covers the data it is claimed to cover".

In this work, we achieved both. In the implementing of aspects of Indonesian morphology, we were able to provide a more exact model of the language. Also in our modelling of morphology, with our morphological analyser, we were equipped with tools that allowed us to design an experiment that checked whether the theory that Indonesian conflated all open word classes was valid or not.

We had successfully shown the application of computational linguistic methods to aid in linguistic questions in this study. In particular, we had successfully con-

[1]These are a list of function words, such as determiners, prepositions and other words belonging to closed class categories that are high in frequency.

structed and executed an unsupervised experiment that supported Yoder's (2010) claim against Gil (1994, 2001, 2010) that nouns and verbs are indeed valid classes in Indonesian.

This work also demonstrated the application of deep lexical acquisition to a language that is relatively under-resourced. Although the semantically-based method was more apt to discover semantically-similar words, we can further investigate methods to discover if there is, and if so, the nature of the relationship between the Levin-style semantically-coherent classes and the *kan* classes that we defined.

Bibliography

Alsina, Alex. 1996. *The Role of Argument Structure in Grammar: Evidence from Romance*. Stanford, USA: CSLI Publications.

Andrews, Avery. 1985. The major functions of the noun phrase. In *Language Typology and Syntactic Description*, ed. by Timothy Shopen, volume 1, chapter 3, 132–223. Cambridge, UK: Cambridge University Press, second edition.

Arantzabe-Urruzolak, María Jesús, 2008. *Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala*. Euskal Herriko Unibertsitatea: Euskal Filologia Saila Departamento de Filología Vasca Ph.D. dissertation.

Arka, I Wayan, 1993. Morphological aspects of the -kan causative in Indonesian. Master's thesis, The University of Sydney, Sydney, Australia.

——. 2003. *Balinese morphosyntax: a lexical-functional approach*. The Australian National University, Australia: Pacific Linguistics.

——. 2008. Voice and the syntax of ā/a verbs in Balinese. In *Voice and Grammatical Relations in Austronesian Languages*, ed. by Peter K. Austin and Simon Musgrave, Studies in Constraint-Based Lexicalism, chapter 3, 45–69. Stanford, USA: CSLI Publications.

——, Avery Andrews, Mary Dalrymple, Meladel Mistica, and Jane Simpson. 2009. A linguistic and computational morphosyntactic analysis for the applicative -i in Indonesian. In *Proceedings of LFG09*, ed. by Tracy Holloway King and Miriam Butt, 85–105.

——, and Christopher D. Manning. 1998. Voice and grammatical relations in Indonesian: A new perspective. In *Proceedings of the 1998 International Lexical Functional Grammar Conference*, Stanford, USA. CSLI Publications.

——, and Christopher D. Manning. 2008. Voice and grammatical relations in Indonesian: A new perspective. In *Voice and Grammatical Relations in Austronesian Languages*, ed. by Peter K. Austin and Simon Musgrave, Studies in

Constraint-Based Lexicalism, chapter 3, 45–69. Stanford, USA: CSLI Publications.

ARTSTEIN, RON, and MASSIMO POESIO. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34.555–596.

ASIAN, JELITA, HUGH E. WILLIAMS, and S. M. M. TAHAGHOGHI. 2005. Stemming indonesian. In *Proceedings of the Twenty-eighth Australasian conference on Computer Science - Volume 38*, ACSC '05, 307–314, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

BALDWIN, TIMOTHY. 2005. Bootstrapping deep lexical resources: Resources for courses. In *Proceedings of the 43rd Annual Meeting of the ACL*, 67–75.

——. 2007. Scalable deep linguistic processing: Mind the lexical gap. In *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation (PACLIC21)*, 3–12, Seoul, Korea.

——, and SUÁD AWAB. 2006. Open source corpus analysis tools for Malay. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*, 2212–5, Genoa, Italy.

——, JOHN BEAVERS, EMILY M. BENDER, DAN FLICKINGER, ARA KIM, and STEPHAN OEPEN. 2005. Beauty and the beast: What running a broad-coverage precision grammar over the bnc taught us about the grammar — and the corpus. In *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*, 49–69. Mouton de Gruyter.

——, STEVEN BIRD, and BADEN HUGHES. 2006. Collecting low-density language materials on the web. In *Proceedings of the 12th Australian World Wide Web Conference (AusWeb06)*, Noosa Lakes, Australia.

BAUER, LAURIE, and VALERA SALVADOR HERNANDEZ. 2005. *Approaches to conversion/zero derivation*. Münster, Germany: Waxmann Verlag.

BEESLEY, KENNETH R., and LAURI KARTTUNEN. 2003. *Finite State Morphology*. Stanford, USA: CSLI Publications.

BENDER, EMILY M. 2009. Reweaving a grammar for wambaya. *Linguistic Issues in Language Technology* 1–34.

——, DAN FLICKINGER, and STEPHAN OEPEN. 2011. Grammar Engineering and Linguistic Hypothesis Testing: Computational Support for Complexity in Syntactic Analysis. In *Language from a Cognitive Perspective*, ed. by Emily M. Bender and Jennifer E. Arnold, Stanford, USA. CSLI Publications.

Berg-Kirkpatrick, Taylor, Alexandre B. Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Proceedings of NAACL 2010*, 582–590, Los Angeles, USA.

Biemann, Chris. 2006. Unsupervised part-of-speech tagging employing efficient graph clustering. In *Proceedings of the COLING/ACL Student Research Workshop*, 7–12, Sydney, Australia.

——. 2009. Unsupervised part-of-speech tagging in the large. *Research on Language and Computation* 7.101–135.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3.993–1022.

Blust, Robert. 2002. Notes on the history of 'focus' in Austronesian languages. In *The history and typology of western Austronesian voice systems*, ed. by Fay Wouk and Malcolm Ross, number 518, chapter 3, 63–78. Canberra, Australia: Pacific Linguistics.

Bonial, Claire, Susan Windisch Brown, Jena D. Hwang, Christopher Parisien, Martha Palmer, and Suzanne Stevenson. 2011. Incorporating coercive constructions into a verb lexicon. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, 72–80, Portland USA.

Brants, S., S. Dipper, S. Hansen, W. Lezius, and G. Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, 24–41, Sozopol, Bulgaria.

Brent, Michael R. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Association for Computational Linguistics* 243–262.

Bresnan, Joan. 1995. Lexicality and argument structure. In *Paris Syntax and Semantics Conference*, Paris, France.

——. 2001. *Lexical Functional Syntax*. Massachusetts, USA: Blackwell Publishers.

——, and Jonni M. Kanerva. 1989. Locative Inversion in Chechewa: A Case Study of Factorization in Grammar. *Linguistic Inquiry* 1–50.

——, and Tatiana Nikitina. 2008. The Gradience of the Dative Alternation. In *Reality Exploration and Discovery: Pattern Interaction in Language and Life*, ed. by Linda Uyechi and Lian-Hee Wee, 1–23. CSLI Publications.

——, and Annie Zaenen. 1990. Deep unaccusativity in LFG. In *Grammatical Relations. A Cross-Theoretical Perspective*, ed. by K. Dziwirek, 45–57. Stanford, USA: CSLI Publications.

Briscoe, Ted, and John Carroll. 1993. Generalized probablistic LR parsing of natural language (corpora) with unification-based grammars. *Association for Computational Linguistics* 25–59.

——, and ——. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, 356–363, Washington DC, USA.

Brody, Samuel, and Mirella Lapata. 2009. Bayesian word sense induction. In *+Proc. of the 16th Conference of the EACL (EACL 2009)*, 103–111, Athens, Greece.

Brown, Peter F., Vincent J. Della Pietra, Peter V. Desouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 14.467–479.

Burchardt, A., K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*, Genoa, Italy.

Butt, Miriam, Stefanie Dipper, Anette Frank, and Tracy Holloway King. 1999a. Writing large-scale parallel grammars for English, French, and German. In *Proceedings of the LFG99 Conference*, ed. by Miriam Butt and Tracy Holloway King. CSLI Publications.

——, and Tracy Holloway King. 2003. Grammar writing, testing, and evaluation. In *Handbook for Language Engineers*, ed. by Ali Ahmed Sabry Farghaly, 129–179. CSLI Publications.

——, ——, and John T. Maxwell III. 2003. Complex predicates via restriction. In *Proceedings of the LFG03 Conference*, ed. by Miriam Butt and Tracy Holloway King, 92–104.

——, ——, Maria-Eugenia Nino, and Frederique Segond. 1999b. *A Grammar Writer's Cookbook*. Number 95 in CSLI Lecture Notes. Stanford, USA: CSLI Publications.

Cahill, Aoife, 2004. *Parsing with Automatically Acquired, Wide-Coverage, Robust, Probabilistic LFG Approximations*. Dublin, Ireland: School of Computing dissertation.

——, and Arndt Riester. 2009. Incorporating information status into generation ranking. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, 817–825, Suntec, Singapore.

CARROLL, JOHN, and ALEX C. FANG. 2004. The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *Proceedings of the First International Joint Conference of Natural Language Processing (IJCNLP-04)*, 107–114, Sanya City, China.

CHRISTODOULOPOULOS, CHRISTOS, SHARON GOLDWATER, and MARK STEEDMAN. 2010. Two decades of unsupervised POS induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 575–584, Massachusetts, USA.

CHUNG, SANDRA. 1976. An object-creating rule in Bahasa Indonesia. *Linguistic Inquiry* 7.41–87.

——. 1978. Stem sentences in Indonesian. In *Second International Conference on Austronesian Linguistics Proceedings*, ed. by Stephen A. Wurm and Lois Carrington, C, 335–365, The Australian National University, Australia. Research School of Pacific Linguistics, ANU.

CLARK, ALEXANDER. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the 10th Annual Meeting of the European Association for Computational Linguistics (EACL)*, 59–66, Budapest, Hungary.

COHN, ABIGAIL, and JOHN J. MCCARTHY. 1998. Alignment and parallelism in Indonesian phonology. *University of Massachusetts Amherst: Linguistics Department Faculty Publication* .

COLE, PETER, and MIN-JEONG SON. 2004. The argument structure of verbs with the suffix -kan in Indonesian. *Oceanic Linguistics* 43.339–364.

COPESTAKE, ANN, and DAN FLICKINGER. 2000. An open-source grammar development environment and broad-coverage English grammar using HSPG. In *Proceeding of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, 591–600, Athens, Greece.

CORBETT, GREVILLE G., DUNSTAN BROWN, and NICHOLAS EVANS. 2002. Morphology, typology, computation. In *Morphology 2000 : Selected Papers From the 9th Morphology Meeting, Vienna, 24-28 February 2000*, ed. by Sabrina Bendjaballah. Amsterdam, Netherlands: John Benjamins Publishing Company.

CROFT, WILLIAM. 2000. Parts of speech as language universals and as language-particular categories. In *Approaches to the typology of word classes*, ed. by P. Vogel and Bernard Comrie, 65–102, Berlin, Germany. Mouton de Gruyter.

——. 2003. *Typology and Universals*. Cambridge, UK: Cambridge University Press.

CROUCH, RICHARD, MARY DALRYMPLE, RONALD KAPLAN, TRACY HOLLOWAY KING, JOHN T. MAXWELL III, and PAULA NEWMAN, 2011. *XLE documentation*. Palo Alto Research Center (PARC), Palo Alto, U.S.A.

——, and TRACY HOLLOWAY KING. 2005. Unifying lexical resources. In *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, 32–37, Saarbrueken, Germany.

DALRYMPLE, MARY. 2001. *Lexical Functional Grammar*, volume 34 of *Syntax and Semantics*. San Diego, USA: Academic Press.

——, RONALD KAPLAN, and TRACY HOLLOWAY KING. 2004. Linguistic generalizations over descriptions. In *Proceedings of LFG04*, 199–208, Christchurch, New Zealand. CSLI Publications.

DARDJOWIDJOJO, SOENJONO. 1971. The meN-, meN-kan, and meN-i verbs in indonesian. *Philippine Journal of Linguistics* 2.71–84.

DIXON, ROBERT MALCOLM WARD. 1994. *Ergativity*. Cambridge Studies in Linguistics. Cambridge, UK: Cambridge University Press.

DONOHUE, MARK. 2004. Voice oppositions without voice morphology. In *Proceedings of AFLA 11, ZAS, Berlin 2004. ZAS Papers in Linguistics*, ed. by Paul Law, volume 34, 73–88. Zentrum für Allgemeine Sprachwissenschaft, Typologie und Universalienforschung.

——. 2010. Covert word classes. In *Part of Speech: Empirical and theoretical advances*, ed. by Umberto Ansaldo, Jan Don, and Roland Pfau, volume 25 of *Benjamins Current Topics*, 87–106. Amsterdam, Germany: John Benjamins Publishing Company.

EVANS, NICHOLAS. 2000. Word classes in the world's languages. In *Morphology: a handbook on inflection and word formation*, ed. by Christian Lehmann Geert Booij and Joachim Mugdan, 708–732, Berlin, Germany. Mouton de Gruyter.

——, and TOSHIKI OSADA. 2005. Mundari: The myth of a language without word classes. *Linguistic Typology* 9.351–390.

FALK, INGRID, CLAIRE GARDENT, and JEAN-CHARLES LAMIREL. 2012. Classifying French verbs using French and English lexical resources. In *Proceedings of the 50th Annual meeting on Association for Computational Linguistics*, 854–863, Jeju, Korea.

FALK, YEHUDA. 2001. *Lexical-Functional Grammar: An introduction to parallel constraint-based syntax*. Stanford, USA: CSLI Publications.

FOLEY, WILLIAM A., 1998. Symmetrical voice systems and precategorality in Philippine languages. Presented at the 3rd LFG Conference, Brisbane, Australia, and republished online.

—— 2008. The place of Philippine languages in a typology of voice systems. In *Voice and Grammatical Relations in Austronesian Languages*, chapter 2, 22–44. CSLI Publications.

FORST, MARTIN, 2007. *Disambiguation for a Linguistically Precise German Parser*. University of Stuttgart: Philosophisch-Historischen Fakultät Ph.D. dissertation.

FOX, EMILY M., ERIK B. SUDDERTH, MICHAEL I. JORDAN, and ALAN. 2009. The sticky HDP-HMM: Bayesian nonparametric Hidden Markov Models with persistent. Technical report, MIT Laboratory for Information and Decision Systems.

FRANCIS, W. NELSON, and HENRY KUČERA, 1979. *Brown Corpus Manual: A Standard Corpus of Present-Day Edited American English, for use with Digital Computers.*. Brown University, Providence, U.S.A.

GIL, DAVID. 1994. The structure of riau indonesian. *Nordic Journal of Linguistics* 17.179–200.

——. 2001. Escaping eurocentrism: fieldwork as a process of unlearning. In *Linguistic Fieldwork*, ed. by Paul Newman and Martha Ratliff, chapter 5. Cambridge, UK: Cambridge University Press.

——. 2005. Word order without syntactic categories: How riau indonesian does it. In *Verb first: On the Syntax of Verb-initial Languages*, ed. by Andrew Carnie, Heide Harley, and Sheila Ann Dooley, 243–263. John Benjamins Publishing Company.

——. 2009. Austronesian nominalism and the thinginess illusion. *Theoretical Linguistics* 35.95–114.

——. 2010. The acquisition of syntactic categories in Jakarta Indonesian. In *Part of Speech: Empirical and theoretical advances*. John Benjamins Publishing Company.

GIRJU, ROXANA, PRESLAV NAKOV, VIVI NASTASE, STAN SZPAKOWICZ, PETER TURNEY, and DENIZ YURET. 2009. Classification of semantic relations between nominals. *Language Resources and Evaluation* 43.105–121.

GODDARD, CLIFF, and BERT PEETERS. 2006. The Natural Semantic Metalanguage (NSM) approach. In *Semantic Primes and Universal Grammar: Empirical evidence from the Romance languages*, ed. by Bert Peeters. Amsterdam, Germany: John Benjamins Publishing Company.

GOLDSMITH, JOHN. 2001. Unsupervised learning of the morphology of a natural language. *Association for Computational Linguistics* 27.153–198.

GORDON, RAYMOND. 2005. *Ethnologue: Languages of the World*. Dallas, USA: SIL International.

GRAHAM, YVETTE, 2011. *Deep Syntax in Statistical Machine Translation*. Dublin, Ireland: Dublin City University dissertation.

GRIFFITHS, THOMAS L., MARK STEYVERS, DAVID M. BLEI, and JOSHUA B. TANENBAUM. 2005. Integrating topics and syntax. *Advances in Neural Information Processing Systems* 17.

HASPELMATH, MARTIN. 2001. *Word Classes and Parts of Speech*, 16538–16545. Pergamon.

HIMMELMANN, NIKOLAUS P. 2005. The austronesian languages of asia and madagascar: Typological characteristics. In *The Austronesian Languages of Asia and Madagascar*, ed. by Alexander Adelaar and Nikolaus P. Himmelmann, chapter 5. New York, USA: Routledge.

—— 2008. Lexical categories and voice in tagalog. In *Voice and Grammatical Relations in Austronesian Languages*, ed. by Peter K. Austin and Simon Musgrave, Studies in Constraint-Based Lexicalism, 249–295. Stanford, USA: CSLI Publications.

JACKENDOFF, RAY S. 1972. *Semantic Interpretation in Generative Grammar*. Cambridge, USA: MIT Press.

—— 1996. The proper treatment of measuring out telecity, and perhaps even quantification in English. *Natural Language and Linguistic Theory* 14.305–354.

—— 2002. What's in the lexicon? In *Storage and Computation in the Language Faculty*, ed. by Sieb Nooteboom, Fred Weerman, and Frank Wijnen, chapter 2, 23–58. Dordrecht, The Netherlands: Kluwer Academic Publishers.

—— 2010. *The Parallel Architecture*. Oxford, UK: Oxford University Press.

JAIN, A.K., M.N. MURTY, and P.J. FLYNN. 1999. Data clustering: A review. In *ACM Computing Surveys*, volume 31 3, 264–323. Association for Computing Machinery.

JOANIS, ERIC, SUZANNE STEVENSON, and DAVID JAMES. 2008. A general feature space for automatic verb classification. *Natural Language Engineering* 337–367.

JUKES, ANTHONY. 2012. Voice, valence and focus in Makassarese. *NUSA - Linguistic Studies of Indonesian and Other Languages in Indonesia* 34.

JURGENS, DAVID, and KEITH STEVENS. 2010. HERMIT: Flexible clustering for the SemEval-2 WSI task. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 359–362, Uppsala, Sweden.

KAPLAN, RON, JOHN T. MAXWELL III, TRACY HOLLOWAY KING, and RICHARD CROUCH. 2004. Integrating finite-state technology with deep LFG grammars. In *Proceedings of the Workshop on Combining Shallow and Deep Processing for NLP (ESSLLI)*.

KAPLAN, RONALD, and JOAN BRESNAN. 1982. Lexical-functional grammar: A formal system for grammatical representation. In *The Mentral Representation of Grammatical Relations*, ed. by Joan Bresnan, 173–281, Cambridge, USA. MIT Press.

——, and JÜRGEN WEDEKIND. 1993. Restriction and correspondence-based translation. In *Proceedings of the 6th conference of EACL*, 193–202, Utrecht, The Netherlands.

KASWANTI, PURWO B. 1997. The direct object in bi-transitive clauses in indonesian. In *Grammatical Relations: a Functional Perspective*, ed. by T. Givón, 233–252. John Benjamins Publishing Company.

KAUFMAN, DANIEL. 2009. Austronesian nominalism and its consequences: A Tagalog case study. *Theoretical Linguistics* 35.1–49.

KAY, MARTIN. 2005. ACL lifetime achievement award: A life of language. *Computational Linguistics* 31.425–438.

KELLER, FRANK, 2001. *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. The University of Edinburgh dissertation.

KIBORT, ANNA, 2004. *Passive and passive-like constructions in English and Polish*. Cambridge, UK: University of Cambridge dissertation.

KIPPER, KARIN, ANNA KORHONEN, NEVILLE RYANT, and MARTHA PALMER. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation* 42.21–44.

KLEIN, DAN, and CHRISTOPHER D. MANNING. 2004. Corpus-Based Induction of Syntactic Structure: Models of Dependency and Constituency. In *Proceeding of the 42nd Annual Meeting on Association for Computational Linguistics*, Barcelona, Spain.

KOO, TERRY, XAVIER CARRERAS, and MICHAEL COLLINS. 2008. Simple semi-supervised dependency parsing. In *Proceedings of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 595–603, Columbus, USA.

KORHONEN, ANNA. 2010. Automatic lexical classification: bridging research and practice. *Philosophical Transactions of The Royal Society* 368.3621–3632.

KRIDALAKSANA, HARIMURTI. 1998. *Introduction to word formation and word classes in Indonesian*. Jakarta, Indonesia: Fakultas Sastra Universitas Indonesia.

KRISHNAMURTHY, RAMESH. 2008. Corpus-driven lexicography. *International Journal of Lexicography* 21.231–242.

KROEGER, PAUL R. 1993. *Phrase structure and grammatical relations in Tagalog*. Stanford, USA: CSLI Publications.

—— 2007. Morphosyntactic vs. morphosemantic functions of Indonesian -kan. In *Architectures, rules, and preferences: variations on themes*, ed. by Joan Bresnan, Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling, and Christopher D. Manning, CSLI Lecture Notes, 229–251. CSLI Publications.

LAPATA, MIRELLA, and CHRIS BREW. 2004. Verb class disambiguation using informative priors. *Computational Linguistics* 30.45–73.

LARASATI, SEPTINA, VLADISLAV KUBON, and DANIEL ZEMAN. 2011. Indonesian morphology tool (MorphInd): Towards an indonesian corpus. In *Proceedings of the Workshop on systems and Frameworks for Computational Morphology*.

LAU, JEY HAN, PAUL COOK, DIANA MCCARTHY, DAVID NEWMAN, and TIMOTHY BALDWIN. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 591–601, Avignon, France. Assocation for Computational Linguistics.

LEVIN, BETH. 1989. *English Verb Classes and Alternations: A preliminary investigation*. Chicago, USA: The University of Chicago Press.

——. 1993. *Towards a lexical organization of English verbs*. Chicago, USA: Chicago Press.

LI, JIANGUO, and CHRIS BREW. 2008. Which Are the Best Features for Automatic Verb Classification? In *Proceeding of the 42nd Annual Meeting on Association for Computational Linguistics*, 434–442.

MACKAY, DAVID J. C., and LINDA C. BAUMAN PETO. 1995. A hierarchical Dirichlet language model. *Natural Language Engineering* 1.289–307.

MACKINLAY, ANDREW, 2012. *Pushing the Boundaries of Deep Parsing*. The University of Melbourne: Computing and Information Systems Ph.D. dissertation.

——, DAVID MARTINEZ, and TIMOTHY BALDWIN. 2012. Detecting modification of biomedical events using a deep parsing approach. *BMC Medical Informatics and Decision Making* 12.S4.

MACLACHLAN, ANNA E., 1996. *Aspects of Ergativity in Tagalog*. McGill University, Canada: Department of Linguistics dissertation.

MALOUF, ROBERT. 1996. A constructional approach to english verbal gerunds. In *Proceedings of the Twenty-Second Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on The Role of Learnability in Grammatical Theory*, 255–266.

MANANDHAR, SURESH, IOANNIS P. KLAPTAFTIS, DMITRIY DLIGACH, and SAMEER S. PRADHAN. 2010. SemEval-2010 Task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 63–68, Los Angeles, USA. Association for Computational Linguistics.

MANNING, CHRISTOPHER D. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 235–242, Columbus, USA.

—— 1996. *Ergativity: Argument Structure and Grammatical Relations*. Cambridge University Press Dissertations in Linguistics Series. Stanford, USA: CSLI Publications.

——, PARBHAKAR RAGHAVAN, and HINRICH SCHÜTZE. 2008. *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press.

——, and HINRICH SCHÜTZE. 2000. *Foundations of Statistical Natural Language Processing*. MIT Press.

MARCUS, MITCHELL P., BEATRICE SANTORINI, and MARY ANNE MARCINKIEWICZ. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics* 19.313–330.

MAXWELL III, JOHN, 2012. XLE project. `http://www2.parc.com/isl/groups/nltt/xle/`.

——, and RONALD KAPLAN. 1994. The interface between phrasal and functional constraints. *Computational Linguistics* 19.571–590.

MCCARTHY, DIANA. 2006. Lexical acquisition. *Encyclopedia of Language and Linguistics* 61–69.

——, JOHN CARROLL, and JUDITA PREISS. 2003. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics* 639–654.

MCDONALD, RYAN, SLAV PETROV, and KEITH HALL. 2011s. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 62–72, Edinburgh, Scotland.

MENESTRINA, DAVID, STEVEN EUIJONG WHANG, and HECTOR GARCIA-MOLINA. 2010. Evaluating entity resolution results. In *Proceedings of the VLDB Endowment*, volume 3 1, 208–219.

MERLO, PAOLA, and SUZANNE STEVENSON. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics* 27.373–408.

MINTZ, MALCOM W. 2002. *An Indonesian and Malay Grammar for Students*. Perth, Australia: Indonesian/Malay Texts and Resources, second edition.

MISTICA, MELADEL, I WAYAN ARKA, TIMOTHY BALDWIN, and AVERY ANDREWS. 2009. Double double, morphology and trouble: Looking into reduplication in Indonesian. In *Proceedings of the 2009 Australasian Language Technology Workshop (ALTW 2009)*, 44–52, Sydney, Australia.

——, TIMOTHY BALDWIN, and I WAYAN ARKA. 2011. Word classes in Indonesian: A linguistics reality or a convenient fallacy in natural language processing? In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, 293–302, NTU, Singapore.

——, JEY HAN LAU, and TIMOTHY BALDWIN. 2013. Unsupervised Word Class Induction for Under-resourced Languages: A Case Study on Indonesian. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, 685–691, Nagoya, Japan.

MUHADJIR. 1981. *Morphology of Jakarta dialect: Affixation and reduplication by Muhadjir translated by Kay Ikranagara*. Badan Penyelenggara Seri NUSA, Universitas Atma Jaya, Jakarta :.

MUSGRAVE, SIMON, 2001. *Non-subject Arguments in Indonesian*. Melbourne, Australia: The University of Melbourne dissertation.

——. 2008. Introduction. In *Voice and Grammatical Relations in Austronesian Languages*, ed. by Peter K. Austin and Simon Musgrave, Studies in Constraint-Based Lexicalism. Stanford, USA: CSLI Publications.

MYCOCK, LOUISE JANE, 2006. *The Typology of Constituent Questions: A Lexical-Functional Grammar analysis of 'wh'-questions*. University of Manchester: Languages, Linguistics and Cultures Ph.D. dissertation.

NASEEM, TAHIRA, REGINA BARZILAY, and AMIR GLOBERSON. 2012. Selective sharing for multilingual dependency. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 629–637, Jeju, Korea.

NICHOLSON, JEREMY, and TIMOTHY BALDWIN. 2008. Learning count classifier preferences of Malay nouns. In *Proceedings of the 2008 Australasian Language Technology Workshop (ALTW 2008)*, 115–123, Hobart, Australia. Australasian Language Technology Association.

——, and ——. 2009. Web and corpus methods for Malay count classifier prediction. In *Proceedings of the 2009 Conference of the NAACL (HLT-NAACL 2009) (Short paper)*, 69–72.

O'DONOVAN, RUTH, MICHAEL BURKE, AOIFE CAHILL, JOSEF VAN GENABITH, and ANDY WAY. 2005. Large-scale induction and evaluation of lexical resources from the Penn-II and Penn-III treebanks. *Computational Linguistics* 229–365.

OEPEN, STEPHAN, HELGE DYVIK, JAN TORE LONNING, ERIC VELLDAL, DOROTHEE BEERMANN, JOHN CARROLL, DAN FLICKINGER, LARS HELLAN, JANNE BONDI JOHANNESSEN, PAUL MEURER, TORBJORN NORDGÅRD, and VICTORIA ROSÉN. 2004. Towards MRS-based Norwegian–English machine translation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, USA.

PALMER, MARTHA, PAUL KINGSBURY, and DAN GILDEA. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31.71–106.

PARISIEN, CHRIS, and SUZANNE STEVENSON. 2010. Learning verb alternations in a usage-based bayesian model. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, Portland, USA.

——, and ——. 2011. Generalizing between form and meaning using learned verb classes. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 2024–2029, Boston, USA.

PAWLEY, ANDY. 2010. Helter skelter and nugl nagl: English and kalam rhyming jingles in psychic unity of mankind. In *A mosaic of languages and cultures: Studies celebrating the career of Karl J. Franklin*, 273–293. Texas, USA: SIL International.

PETERSON, DAVID A. 2007. *Applicative Constructions*. Oxford University Press.

PISCELDO, FEMPHY, RAHMAD MAHENDRA, RULI MANURUNG, and I WAYAN ARKA. 2008. A two-level morphological analyser for Indonesian. In *Proceedings of the Australasian Language Technology Association Workshop*, volume 6, 88–96.

——, RULI MANURUNG, and MIRNA ADRIANI. 2009. Probablilistic part-of-speech tagging for Bahasa Indonesia. In *The Third International MALINDO Workshop*, Singapore, Singapore.

PORTERFIELD, LESLIE, and VENEERTA SRIVASTAV. 1988. (In)definiteness in the absence of articles: Evidence from Hindi and Indonesian. In *Proceedings of the Seventh West Coast Conference on Formal Linguistics*, ed. by Hagit Borer, 265–276.

POSTMAN, WHITNEY, 2002. *Thematic role assignment in Indonesian: A case study of agrammatic aphasia*. Ithaca, USA: Department of Linguistics dissertation.

QUINN, GEORGE. 2001. *The learner's dictionary of today's Indonesian*. Allen & Unwin, St Leonards, N.S.W. :.

RAMOS, TERESITA V., and MARIA LOURDES S. BAUTISTA. 1986. *Handbook of Tagalog Verbs*, volume 301. Honolulu: USA: University Press of Hawaii.

RIZA, HAMMAM, T. SYARFINA, ARIEF SARTONO, EGGI GITHA NAGARA, AWAL SUBANDAR, DARWITO, BUDIONO, HENKY MULYADI, and ADIANSYA PRAETYA. 2008. Initial research report on corpus design, collection and cleaning tools. Technical report, Agency for the Assessment and Application of Technology (BPPT), Jakarta, Indonesia.

ROSÉN, VICTORIA, and ANNIE ZAENEN. 1999. Grammar writing in LFG – Introduction. In *Proceedings of the LFG99 Conference*. CSLI Publications.

SASSE, HANS-JÜRGEN. 2001. Scales between nouniness and verbiness. In *Language Typology and Language Universals*, ed. by Martin Haspelmath, , Ekkard König, Wulf Oesterreicher, and Wolfgang Raible, 495–509. Berlin, Germany: Walter d Gruyter.

SCHACHTER, PAUL. 1985. Part-of-speech systems. In *Language Typology and Syntactic Description*, ed. by Timothy Shopen, volume 1, 3–61. Cambridge University Press.

SCHULTE IM WALDE, SABINE. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics* 32.159–194.

——. 2009. The Induction of Verb Frames and Verb Classes from Corpora. In *Corpus Linguistics. An International Handbook.*, ed. by Anke Lüdeling and Merja Kytö, volume 2 of *Handbooks of Linguistics and Communication Science*, chapter 44, 952–971. Berlin: Mouton de Gruyter.

SMITH, NOAH A. 2011. *Linguistic Structure Prediction.* Synthesis Lectures on Human Language Technologies. Morgan and Claypool Publishers.

SNEDDON, JAMES NEIL. 1996. *Indonesian reference grammar.* St Leonards, Australia: Allen & Unwin.

——, ALEXANDER ADELAAR, DWI NOVERINI DJENAR, and MICHAEL C. EWING. 2010. *Indonesian reference grammar.* St Leonards, Australia: Allen & Unwin, 2nd edition.

SØGAARD, ANDERS. 2012. Unsupervised dependency parsing without training. *Natural Language Engineering* 18.187–203.

SON, MIN-JEONG, and PETER COLE. 2008. An event-based account of -kan construction in standard Indonesian. *Language* 84.120–160.

STEVENS, ALAN M., and A. ED. SCHMIDGALL-TELLINGS. 2004. *Kamus Lengkap Indonesia-Inggris.* Jakarta, Indonesia: PT Mizan Pustaka.

SUGIONO, DENDY (ed.) 2008. *Kamus Besar Bahasa Indonesia - Pusat Bahasa.* Departemen Pendidkan Nasional, Jakarta, Indonesia: PT Gramedia Pustaka Utama, edisi keempat edition.

SULGAR, SEBASTIAN, MIRIAM BUTT, TRACY HOLLOWAY KING, PAUL MEURER, TIBOR LACZKÓ, GYORGY RÁKOSI, CHEIKH BAMBA DIONE, HELGE DYVIK, VICTORIA ROSÉN, KOENRAAD DE SMEDT, AGNIESZKA PATEJUK, OZLEM CETINOGLU, WAYAN ARKA, and MELADEL MISTICA. 2013. Pargrambank: The pargram parallel treebank. In *Proceeding of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria.

SUN, LIN, 2012. *Automatic induction of verb clases using clustering.* Cambridge, UK: University of Cambridge dissertation.

——, and ANNA KORHONEN. 2011. Hierarchical verb clustering using graph factorization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1023–1033, Edinburgh, Scotland.

——, ——, and YUVAL KRYMOLOWSKI. 2008. Verb class discovery from rich syntactic data. In *Proceedings of the 9th International CICLing Conference*, 16–27, Haifa, Israel.

——, ——, THIERRY POIBEAU, and CÉDRIC MESSIANT. 2010. Investigating the cross-linguistic potential of VerbNet-style classification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 1056–1064, Beijing, China. Association for Computational Linguistics.

TÄCKSTRÖM, OSCAR, RYAN MCDONALD, and JAKOB USZKOREIT. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 12th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 477–487, Montreal, Canada.

TAN, PANG-NING, MICHAEL STEINBACH, and VIPIN KUMAR. 2006. *Introduction to Data Mining*. Boston, USA: Addison Wesley.

TEH, YEE WHYE, MICHAEL I. JORDAN, MATTHEW J. BEAL, and DAVID M. BLEI. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101.1566–1581.

TENG, STACY FANG-CHING. 2008. *A Reference Grammar of Puyuma*, volume PL 595. The Australian National University: Pacific Linguistics.

ULINIANSYAH, MOHAMMAD TEDUH, SHUN ISHIZAKI, and KIYOKO UCHIYAMA. 2002. Indonesian morphological parser with minimum connectivity cost to solve ambiguities. In *Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence*, PRICAI '02, p. 606, London, UK. Springer-Verlag.

VAMARASI, MARIT KANA. 1999. *Grammatical relations in Bahasa Indonesia*, volume 93 of *Series D*. The Australian National University, Australia: Pacific Linguistics.

VAN VALIN, ROBERT D., and RANDY J. LAPOLLA. 1997. *Syntax: Structure, Meaning and Function*. Cambridge, UK: Cambridge University Press.

VOSKUIL, JAN. 2000. Indonesian voice and A-bar movement. In *Formal Issues in Austronesian Linguistics*, ed. by Ileana M. Paul, Vivianne Phillips, and Lisa Travis, volume 49 of *Natural Language and Linguistic Theory*, 195–214. Kluwer Academic Publishers.

WALLACH, HANNA M. 2006. Topic modelling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, 977–984, Pittsburgh, USA.

WALSH, MICHAEL. 1996. Vouns and nerbs: A category squish in Murrinh Patha (Northern Australia). In *Studies in Kimberley languages in honour of Howard H. Coate*, 227–252. Munich, Germany: Lincom Europa.

WICAKSONO, ALFAN FARIZKI, and AYU PURWARIANTI. 2010. HMM based part-of-speech tagger for Bahasa Indonesia. In *Proceedings of the Third MALINDO Workshop*, Jakarta, Indonesia.

WIERZBICKA, ANNA. 1996. *Semantics: Primes and Universals*. Oxford, UK: Oxford University Press.

WITTEN, IAN H., and EIBE FRANK. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Data Management Systems. Burlington, USA: Morgan Kaufmann Publishers Inc.

XIA, FEI, and WILLIAM D. LEWIS. 2009. Applying NLP Technologies to the Collection and Enrichment of Language Data on the Web to Aid Linguistic Research. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Science, Humanities, and Education (LaTeCH – SHELT & R)*, 51–59, Athens, Greece.

YEH, ALEXANDER. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th of Conference on Computational Linguistics COLING*, 947–953.

YODER, BRENDON. 2010. Syntactic underspecification in Riau Indonesian. *Working Papers of the Summer Institute of Linguistics* 50.

Appendix A

ParGram Development Data

This data is the testsuite for development.

A.1 September 2009

```
# sentences for the fall 2009 f-structure comparison
###################################
# a simple transitive for a warm-up

# 1. Mary ate the cake.

Mary makan kue itu. (1 0.020 56)

###################################
# NN compounds

# productive
# 2. the book cover

DP: kulit buku itu (2 0.020 31)

# lexicalized
# 3. ice cream

NP: es krim (2 0.020 20)

# various modifiers

# adj modifying N2 (or N1+N2)
# 4. the broken wine bottle

NP: botol anggur yang pecah (2 0.030 42)

# adj modifying N1
# 5. the red wine bottle

NP: botol anggur merah (1 0.020 25)

# adj modifying N2
# 6. the tractor electrical switch
```

```
NP: saklar listrik traktor itu (3 0.020 45)

# NNN
# 7. the oil can manufacturer

NP: pabrik kaleng minyak itu (3 0.020 45)

# 8. the steel oil can

NP: kaleng minyak dari baja itu (4 0.020 46)

# proper noun N1
# 9. the California coast

NP: pantai California (1 0.020 11)

# coordination of N1
# 10. romance and mystery novels

NP: novel misteri dan roman (3 0.010 39)

# punctuation options
# 11. diesel-engine repair

NP: bengkel mesin-disel (1 0.020 26)


##############################################
# pronoun types at the behest of last meeting

# personal pronoun and reflexive pronoun
# 12. He hit himself.

Dia memukul dirinya. (2 0.020 28)

# demonstrative pronoun and reciprocal pronoun
# 13. Those resemble each other.

Mereka saling mirip. (1 0.020 19)

# interrogative pronoun
# 14. Who left?

Siapa yang pergi? (1 0.010 35)

# relative pronoun
# 15. the boy who left

DP: anak yang pergi itu (1 0.020 29)

# expletive pronoun
# 16. It is raining.

Sedang hujan. (2 0.010 22)

# free pronoun
# 17. whoever left

NP: siapapun yang pergi (1 0.010 17)

################################################
# simple ADJUNCT-QT at the behest of last meeting
```

```
# 18. "I hopped," said the girl.

# NOT DONE at time of ParGram 2009
```

A.2 March 2010

```
#### ParGram March 2010
#########################

#1  The monkey will go.

Kera itu akan pergi. (1 0.020 45)

#2  The monkey is laughing.

Kera itu sedang tertawa. (1 0.020 47)

#3  The monkey will be afraid.

Kera itu akan takut. (1 0.010 42)

#4  The monkey was eating a banana.

Kera itu sedang makan pisang. (1 0.020 73)

#5 a.  The monkey gave a bone to the dog.

Kera itu memberikan anjing itu tulang. (3 0.040 192)

#5 b. The monkey gave a bone to the dog.

# Kera itu memberi tulang kepada anjing itu.

#6  The monkey taught the dog tricks.

Kera itu mengajari anjing itu tipu-daya. (10 0.090 196)

#7  a girl suitable for the job

DP: gadis yang pantas untuk pekerjaan itu (4 0.030 55)

#8  a suitable girl for the job

DP: gadis yang pantas untuk pekerjaan itu (4 0.010 55)

#9 a. the eaten apple (deliberate eating)

DP: apel yang sudah dimakan itu (1 0.020 34)

#9 b. the eaten apple (deliberate eating)

#apel yang termakan itu

#10 the tea drinking woman

DP: wanita peminum teh itu (6 0.010 50)
```

```
#11 the woman drinking tea

DP: wanita yang minum teh itu (2 0.010 51)

#12 a. the world's fastest car (possessive)

DP: mobil tercepat dunia itu (2 0.020 38)

#12 b. the world's fastest car (in/of the world)

# DP: mobil tercepat di dunia itu

#13  a car faster than light

NP: mobil yang lebih cepat dari cahaya (6 0.020 52)

#14  John fights with Michael.

John berkelahi dengan Michael. (1 0.020 49)

#15  John trusts in God.

John percaya pada Tuhan. (1 0.010 29)

#16  John resigns from the job.

John mundur dari pekerjaan. (8 0.020 45)

#17  The monkey made the dog go.

Kera itu membuat anjing itu pergi. (3 0.030 140)

#18  The monkey made the dog laugh.

Kera itu membuat anjing itu tertawa. (3 0.030 137)

#19  The monkey made the dog eat the food.

Kera itu memaksa anjing itu makan makanan itu. (6 0.050 229)

#20  The monkey made the dog teach tricks to the cat.

Kera itu memaksa anjing itu mengajari kucing itu tipu-daya. (48 0.650 397)

#21  The dog made the monkey pinch the cat.

Anjing itu memaksa kera itu mencubit kucing itu. (4 0.030 178)

#22  The monkey was made to pinch the cat by the dog.

Kera itu dipaksa mencubit kucing itu oleh anjing itu. (1 0.020 128)

#23  The dog was made to eat the food by the monkey.

Anjing itu dipaksa makan makanan itu oleh kera itu. (1 0.030 120)
```

A.3 October 2010

```
#### ParGram October 2010
##########################

# 1. The thirsty crow

NP: burung gagak yang haus (1 0.030 15)

# 2. A crow was very thirsty.

Ada burung gagak yang haus sekali. (1 0.020 35)

# 3. He came out in search of water.

Dia keluar untuk mencari air. (1 0.020 64)

# 4. He saw a pitcher of water under a tree.

Di bawah pohon, dia melihat kendi yang berisi air. (1 0.040 141)

# 5. There was only a little water in the pitcher.

Hanya ada air sedikit saja di dalam kendi itu. (8 0.030 161)

# 6. The monkey taught the dog tricks.

Paruhnya tidak mencapai airnya. (4 0.030 49)

# 7. And so he could not drink it.

Dengan demikian, dia tidak bisa minum. (6 0.010 49)

# 8. He saw stones on the ground
#Dia melihat banyak batu-kerikil yang tergeletak di tanah

Dia melihat banyak batu-kerikil di tanah. (2 0.030 116)

# 9. He picked up some pebbles from the ground, and put them in the water.

Dia mengangkat beberapa batu-kerikil dan menaruhnya di dalam air. (1 0.030 139)

# 10. The water rose higher.

Airnya semakin tinggi. (2 0.010 29)

# 11. The crow drank some water and flew away.

Burung gagak itu minum air sedikit dan terbang. (12 0.070 282)

#Burung gagak itu minum air sedikit.
```

A.4 October 2011

```
#### ParGram October 2011
##########################

# 1. The ball flew through a broken door and landed in the cellar.
```

Bola itu melayang lewat pintu yang pecah dan mendarat di gudang bawah tanah.

2. One of the children, the 14-year-old daughter of the concierge, hobbled down after it.

Salah seorang anak, putri dari pramupintu yang berumur 14 tahun, mengejarkannya turun dengan terpincang-pincang.

3. A tram had cut off her legs, and she was happy enough if she could pick up the balls
after her friends.

Kakinya terpotong terlindas oleh tram, dan dia cukup senang jika dia bisa memungut bola itu setelah temannya.

4. The cellar was rather dark, but she thought she could see something stirring in the corner.

Gudang bawah tanah itu agak gelap, tetapi dia pikir dia bisa melihat sesuatu yang bergerak di sudut.

5. 'What are you doing here, little kitty?', the wooden-legged girl called out.
'Apa yang kau lakukan di sini, kucing kecil?', gadis yang berkaki-kayu itu berteriak.
=> Vocatives not done
=> Direct quotes "" not done

Apa yang kau lakukan di sini?

6. She then picked up the ball, and hurried off as fast as possible.'

Dia kemudian memungut bola itu, dan bergegas pergi secepatnya.

7. The old ugly and foul-smelling rat, which had been taken for a kitten, was stunned.

Tikus besar yang jelek dan berbau itu, yang telah dikira sebagai anak kucing, terkejut.

8. No one had ever talked to it like that before.

Tidak ada orang yang berbicara seperti itu sebelumnya.

9. if only it had been born a kitten, or better yet, the lame daughter of the concierge

jika dia saja lahir sebagai anak kucing atau, lebih baik lagi, anak dari pramupintu

10. But this thought was so very beautiful, the rat couldn't even imagine it in earnest.

Tetapi pikiran itu begitu indahnya, tikus itu bahkan tidak bisa membayangkannya dengan sungguh-sungguh.

A.5 July 2012

```
# == ParGram Parallel Corpus
# ==========================

### = Basic sentence types

### - Declaratives

# 1. The driver starts the tractor.

Sopir itu menghidupkan traktor itu. (5 0.030 84)

# 2. The tractor is red.

Traktor itu merah. (1 0.020 39)
```

```
### - Interrogatives

# 3. What did the farmer see?

Apa yang dilihat petani itu? (1 0.020 65)

# 4. Did the farmer sell his tractor?

Apakah petani itu sudah menjual traktornya? (4 0.040 87)

### - Imperatives

# 5. Push the button.

Pencet tombol itu. (1 0.020 43)

# 6. Don't push the button.

Jangan pencet tombol itu. (1 0.020 46)

### - Transitivity (7. Di; 8. Trans; 9. Intrans)

# 7. The farmer gave his neighbour an old tractor.

Petani itu memberikan tetangganya sebuah traktor tua. (14 0.090 330)

# 8. The farmer cut the tree down.

Petani itu memotong pohon itu. (1 0.020 69)

# 9. The farmer groaned. (i.e. not a quotative: *groaned I'm poor old man)

Petani itu mengaduh. (1 0.010 35)

### - Passives and traditional voice -> also see Reflexives

# 10. My neighbour was given an old tractor by the farmer.

Tetanggaku diberikan sebuah traktor tua oleh petani itu. (20 0.150 261)

# 11. The tree was cut down yesterday.

Pohon itu dipotong kemarin. (1 0.020 47)

# 12. The tree had been cut down. (i.e. obviously because it's not there)

Pohon itu sudah terpotong. (1 0.020 47)

# 13. The tractor starts with a shudder.

Traktor bergetar saat dihidupkan. (10 0.020 111)

### - Tracy's favourite

# 14. The tractor appeared.

Traktor muncul. (1 0.020 34)

### = Embedded clauses

# - Subcategorised declaratives
```

```
# 15. The boy knows (that) the tractor is red.

Anak laki-laki itu tahu bahwa traktor merah. (12 0.030 169)

# 16. The child thinks he started the tractor.

Anak itu pikir dia menghidupkan traktor itu. (10 0.030 142)

# - Subcategorised interrogatives

# 17. The farmer knows who started the tractor.

Petani itu tahu siapa yang menghidupkan traktor itu. (20 0.030 229)

# 18. The child wondered whether the button had been pushed.

#Anak itu bertanya-tanya kalau tombol sudah dibeli.

# - Relative clauses and free relatives

# 19. The tractor that the farmer bought is red.

#Traktor yang dibeli petani itu berwarna merah.

# 20. The man who bought the tractor left.

#Orang yang membeli traktor itu sudah pergi.

# 21. The store the farmer bought the tractor from closed. (i.e. no longer trading)

#Toko dari mana petani membeli traktor itu ditutup.

# 22. Whoever bought this tractor is a lucky person.

#Siapa yang membeli traktor itu adalah orang yang ketiban pulung.

### = Causative/Permissive

# 23. The farmer made his son clean the tractor.

#Petani itu memaksa anaknya membersihkan traktor itu.

# 24. The farmer made his son leave.

# threw his child out

# Petani itu mengeluarkan anaknya.

# made his child leave

#Petani itu memaksa anaknya pergi.

# 25. The farmer made her son buy the tractor.

#Petani itu memaksa anaknya membeli traktor.

# 26. The farmer let her son buy the tractor.

#Petani itu membebaskan anaknya membeli traktor.

### = Benefactive/Dative Alternation

# 27. The farmer bought his son a tractor.
```

```
#Petani itu membelikan anaknya traktor.
# 28. The woman bought the tractor for her husband.
#Perempuan itu membeli traktor itu untuk suaminya.
# = Reflexive/Reciprocal
# 29. The lovers danced until dawn.   (i.e. with each other)
#Dua kekasih itu berdansa sampai dinihari.
# 30. The boy bathed in the river.   (i.e. himself)
#Anak laki-laki mandi di dalam sungai.
# 31. The teacher read to himself aloud.
#Guru itu membacakan dirinya keras-keras.
# 32. The brothers bought the tractor for each other.
#Orang yang bersaudara itu saling membelikan traktor.
### = Copula, and non-verbal PREDs
# 33. My sister is a great teacher.
#Saudara perempuan saya adalah guru yang baik.
# 34. The child is in the house.
#Anak itu di dalam rumah.
# 35. The children are happy.
#Anak-anak senang.
### = Expletive pronouns
# 36. It is raining.
#Sedang hujan.
# 37. There is a problem with the tractor.
#Ada masalah tentang traktor.
### = NN compounds
### - Productive
# 38. The book cover depicted a tractor.
#Kulit buku menggambarkan traktor.
### - Lexicalised
# 39. Let's get ice-cream.
#Mari kita makan es krim.
```

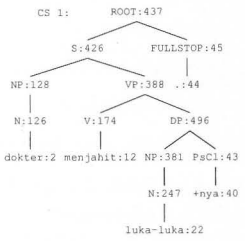
```
### - Adj modifying N2 (or N1+N2)

# 40. The boy swept up the broken wine bottle.

#Anak itu menyapu botol anggur pecah.

### - Adj modifying N2

# 41. The red wine bottle broke.

#Botol anggur merah itu dipecah.

### = Adjectives

### - Stacked adjectives

# 42. There are great green globs of greasy grimy gopher guts.

#Ada timbunan lemak yang besar dan hijau dari usus binatang gopher.

### - Arguments of an adjective

# 43. They are proud of their daughter.

#Mereka bangga dengan anak mereka.

### - Comparative

# 44. My tractor is faster than your sports car.

#Traktor saya lebih cepat dari mobil sport anda.

### - Superlative

# 45. My tractor is the fastest vehicle in the county.

#Traktor saya adalah kandaran yang tercepat di kecamatan.

### - Deverbal

# 46. The barking dog woke the neighbours.

#Anjing yang menggonggong itu membangunkan orang-orang sebelah.

# 47. The tea drinking woman admired her new purchase from eBay. (No, they are not sponsoring us!)

#Wanita peminum teh itu mengagumi pembeliannya dari eBay.

### = XCOMPS

# 48. The farmer wants to buy a tractor.

#Petani itu ingin membeli traktor.

# 49. The farmer's daughter promised to repair the tractor.

#Anak perempuan petani itu berjanji akan memperbaiki traktor itu.

# 50. The farmer persuaded his wife to buy a new tractor.

#Petani membujuk istrinya untuk membeli traktor.

### == TO DO:
```

```
### - Coordination

#  Some suggestions:

#  The farmer started the tractor and drove off.

#Petani itu menghidupkan traktor itu lalu mengendarainya.

#  The farmer and his wife bought a new tractor.

#Petani itu dan istrinya membeli traktor baru.

#  The farmer grows carrots, but she doesn't grow celery.

#Petani itu menanam wartel tetapi dia tidak menanam seladri.

#  Also think about coordination of AdjP, A, N, V

#- Pronouns
#- More more more

#=============================================================================================
```

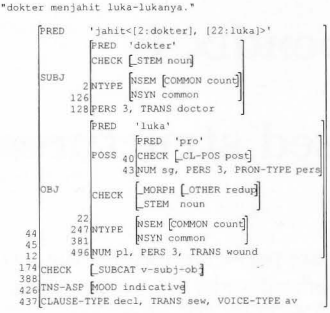
Appendix B

Parsed structures for -kan

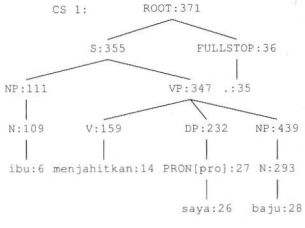
These are parses representing the 5 types summarised in Table 4.1 in Chapter 4. The *kan*-affixed verbs are parsed alongside their non-*kan* counterparts in Figures B.1 to B.2 to show the effects of each defined type in Figure 4.9. The c-structure and f-structure for the non-*kan* verb construction are represented in (a) and (b), respectively, and the c- and f-structures for the *kan*-affixed verbs are shown in (c) and (d).



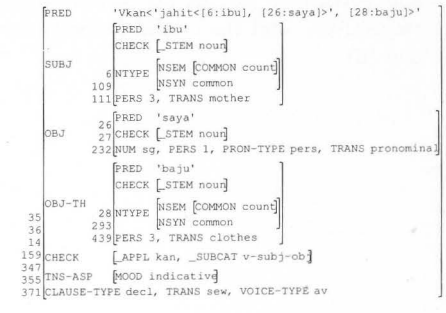
a.



b.



c.

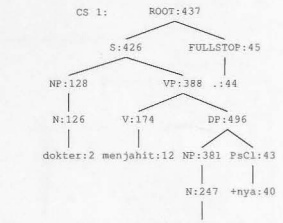


d.

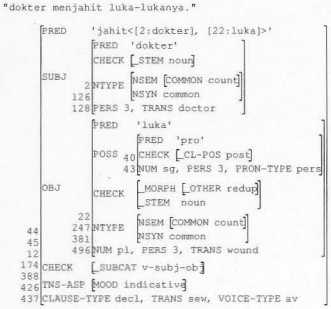
a&b. *dokter menjahit luka-lukanya*
doctor AV-buy wound-wound=3.sg
"the doctor sewed his/her wound"

c&d. *Ibu menjahitkan saya baju.*
mother AV=sew-KAN 1.sg shirt
"Mother sewed me a shirt."

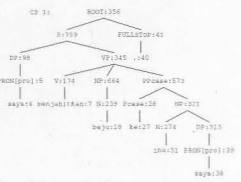
Figure B.1: c-stucture and f-structure for Type 1



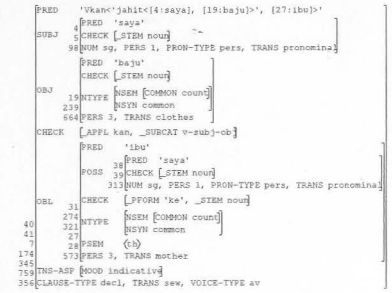
a.



b.



c.

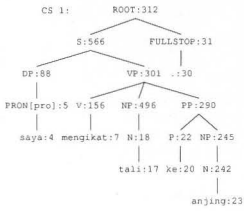


d.

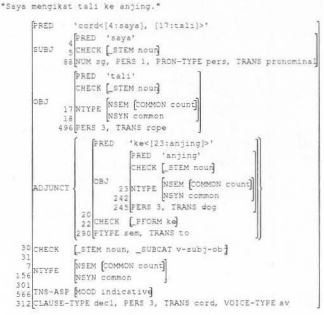
a&b. *dokter menjahit luka-lukanya*
doctor AV-buy wound-wound=3.sg
"the doctor sewed his/her wound"

c&d. *Saya menjahitkan baju ke ibu saya.*
1.sg AV-sew-KAN shirt to mother 1.sg
"I had my shirt sewn by my mum."

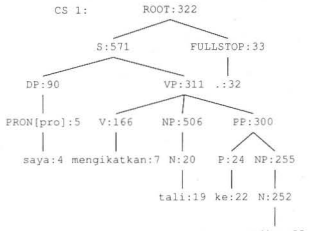
Figure B.2: c-stucture and f-structure for Type 2



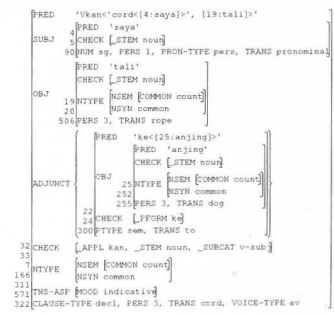
a.



b.



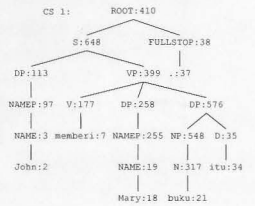
c.



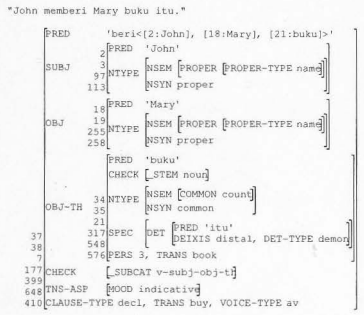
d.

For: *Dia* *mengikat(-kan)* *tali* *itu* *ke* *anjing*
3.sg AV-tie-KAN rope that to dog
"S/he ties the rope to the dog."

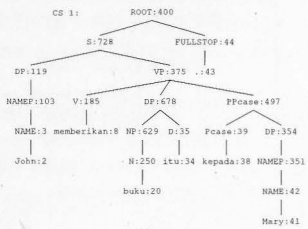
Figure B.3: c-stucture and f-structure for Type 3



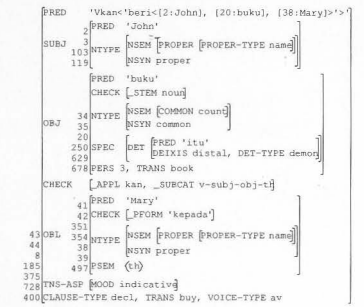
a.



b.



c.



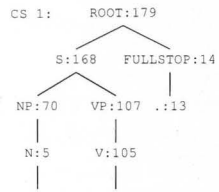
d.

a&b. *John memberi Mary buku itu.*
J AV-give M book that
"John gave Mary the book."

c&d. *John memberikan buku itu kepada Mary.*
J AV-give-KANbook that to M
"John gave the book to Mary."

Figure B.4: c-stucture and f-structure for Type 4

a.

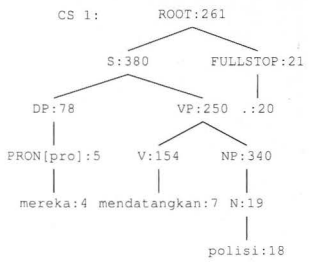


b.

"Polisi datang."

```
PRED     'datang<[4:polisi]>'
SUBJ     PRED   'polisi'
         CHECK  [_STEM noun]
       4 NTYPE  [NSEM [COMMON count]
       5        [NSYN common
      13 PERS 3, TRANS police
      14
       6 CHECK  [_MORPH  [_PRE none]
     105        [_SUBCAT v-subj]
     107
     168 TNS-ASP [MOOD indicative]
     179 CLAUSE-TYPE decl, TRANS come, VOICE-TYPE bv
```

c.



d.

"Mereka mendatangkan polisi."

```
PRED     'Vkan<'datang<[4:mereka]>', [18:polisi]>'
SUBJ   4 PRED   'mereka'
       5 CHECK  [_STEM noun]
      78 NUM pl, PERS 3, PRON-TYPE pers, TRANS pronomina
OBJ      PRED   'polisi'
         CHECK  [_STEM noun]
      18 NTYPE  [NSEM [COMMON count]
      19        [NSYN common
     340 PERS 3, TRANS police
  20 154 CHECK  [_APPL kan, _SUBCAT v-subj]
  21 250
   7 380 TNS-ASP [MOOD indicative]
     261 CLAUSE-TYPE decl, TRANS come, VOICE-TYPE av
```

a&b. *Polisi datang.*
police come/arrive
"The police arrived/came."

c&d. *Mereka mendatangkan polisi.*
3.pl AV-come-KAN police
"The called for/made the police come"

Figure B.5: c-stucture and f-structure for Type 5

Appendix C

100 Stems

C.1 Verb Stems

Table C.1 shows the classification for the verb stems, where the VERB FRAME column indicates possible subcategorisation frames for the MEN+stem and the MEN+stem+KAN morphological patterns, and DECOMPOSITION refers to its semantic description relative to the stem. Also note that '{...}' indicates optionality.

Most verbs were easily classified, but the following paragraphs are notes made during the manual classification where clarification may have been needed.

Notes on Verb Type 4 This is a CAUSE-TO-HAPPEN frame, and it may seem like *masuk* "enter", *hadir* "be present", *lulus* "go through" should be CAUSE-TO-DO verbs, when affixed with *-kan*, but *memasukkan* (stem = *masuk* "enter") means "to put something inside something else", not "make someone enter". There is no volitionality on the part of the *causee*. If a person was the *causee* then the *causer* would have to physically pick up the person and put him/her inside something. The verb *menghadirkan* (stem = *hadir* "be present") means "to summon", for example "summon to court". Also the verb *meluluskan* (stem = *lulus* "go through") means "make to go/allow though".

Notes on Verb Type 7 These verbs in the MEN+stem+KAN frame subcategorie for a VP, which can have an optional 'untuk' before the VP, but does not change the overall meaning.

C.2 Adjective Stems

For adjectives, we needed a way to differentiate between attributing a quality to something, for thinking that something is *remeh* "unimportant", and also feel-

MORPHOLOGY	VERB FRAME	DECOMPOSITION
Type 1: *acuh* "to heed", *terjemah* "translate", *mandi* "bathe"		
MEN+V_1	–	–
MEN+V_1+KAN	<NP_a, NP_b >	DO_{to}([NP_a], [V_1 TO([NP])])
Type 2: *bawa* "carry", *beri* "give"		
MEN+V_2	<NP_a, NP_b > {PP_c}	DO_{to}([NP], [V_2 TO([NP]) {[$_{path}$ P_c([NP_c])]})
MEN+V_2+KAN	<NP_a, NP_b > {PP_c}	DO_{to}([NP], [V_2 TO([NP]) {[$_{path}$ P_c([NP_c])]})
	<NP_a, NP_b, NP_c >	DO_{for}([NP_a], [V_2 TO([NP_c]) FOR([NP_b])])
Type 3: *dengar* "hear", *kenang* "think of"		
MEN+V_3	<NP_a, NP_b >	$HAPPEN_{to}$([NP_b], [V_3 TO([NP_a])])
MEN+V_3+KAN	<NP_a, NP_b >	DO_{to}([NP_a], [V_3 TO([NP_b])])
Type 4: *hidup* "be alive", *jatuh* "fall", *mati* "die", *tewas* "perish", *pusing* "to concern oneself", *minggir* "put aside", *masuk* "enter", *hadir* "be present", *lulus* "go through"		
MEN+V_4	–	–
MEN+V_4+KAN	<NP_a, NP_b >	CAUSE([NP_a], [$HAPPEN_{to}$([V_4 TO([NP_b])])])
Type 5: *serah* "surrender", *singkir* "get out of way", *susup* "duck down"		
ME+V_5+N	<NP_a >	DO([NP_a], [V_5])
MEN+V_5+KAN	<NP_a, NP_b >	CAUSE([NP_a], [$HAPPEN_{to}$([V_5 TO([NP_b])])])
Type 6: *bangun* "form/take shape", *pecah* "break"		
MEN+V_6	<NP_a, NP_b >	CAUSE([NP_a], [$HAPPEN_{to}$([V_6 TO([NP_b])])])
MEN+V_6+KAN	<NP_a, NP_b >	CAUSE([NP_a], [$HAPPEN_{to}$([V_6 TO([NP_b])])])
Type 7: *force* "paksa", and also *buat* "make/do"		
MEN+V_7	<NP_a, NP_b > {VP_c}	DO_{to}([NP_a], [V_7 TO([NP_b]) {[DO([NP_b], [VP_c])]})
MEN+V_7+KAN	<NP_a, NP_b, VP_c >	DO_{to}([NP_a], [V_7 TO([NP_b]) [DO([NP_b], [VP_c])]])
	<NP_a, VP_b >	DO_{to}([NP_a], [V_7 [DO_{to}([VP_b])]])
Type 8: *timpa* "hit", *baca* "read"		
MEN+V_8	<NP_a, NP_b >	DO_{to}([NP_a], [V_8 TO([NP])])
MEN+V_8+KAN	<NP_a, NP_b > {PP_c}	DO_{to}([NP_a], [V_8 TO([NP]) {[$_{path}$$P_c$([$NP_c$])] })

Table C.1: Verb Types, where '–' indicates no attested word form in the text collection.

MORPHOLOGY	VERB FRAME	DECOMPOSITION
Type 1: *abadi* "eternal", *asing* "separated", *cemar* "dirty", *cerdas* "intelligent", *goyah* "unstable", *haram* "prohibited", *murni* "pure", *mutakhir* "recent/up-to-date", *padu* "compact/solid", *populer* "popular", *salah* "wrong", *subur* "fruitful", *terang* "clear"		
MEN+A_1	–	–
MEN+A_1+KAN	<NP_a, NP_b >	<NP_a, CAUSE([NP_b], [BE([A_1)]])
Type 2: *biasa* "ordinary/common", *unggul* "excellent/ahead", *berani* "audacious"		
MEN+A_2	–	–
MEN+A_2+KAN	<NP_a, NP_b > {(PP_c)|(VP_d)}	CAUSE([NP_a], [BE([NP_b], [A_1], {[$_{path}P_c$([NP_c])]|[$_{event}VP_d$]})])
Type 3: *cengang* "amazed", *takjub* "surprised"		
MEN+A_3	–	–
MEN+A_3+KAN	<NP_a >	CAUSE([NP_a], [HAPPEN([A_3])])
Type 4: *kecewa* "disappointed", *leceh* "worthless", *remeh* "unimportant", *teguh* "strong", *jengkel* "annoyed"		
MEN+A_4	–	–
MEN+A_4+KAN	<NP_a >	CAUSE([NP_a], [FEEL([NP_a] [A_4])])
Type 5: *lunak* "soft", *lanjut* "protracted"		
MEN+A_5	<NP_a >	BECOME([A_5])
MEN+A_5	<NP_a >	GO([NP_a], TO([$_{state}$BE([A_5])])
MEN+A_5+KAN	<NP_a >	CAUSE([NP_a], [FEEL([NP_a] [A_1])])

Table C.2: Adjective Types

ing *remeh* "unimportant", and so for adjectives we introduce an NSM primitive FEEL (Wierzbicka 1996:119).

Notes on Adjective Type 2 These stems in the MEN+stem+KAN frame subcategorise for a VP, which can have an optional 'untuk' before the VP, but does not change the overall meaning.

Notes on Adjective Type 3 These stems do not often appear unaffixed in the text, and often occur before the relativiser *yang*. There is never an *experiencer* mentioned, just the stimulus, which is why we describe these as HAPPEN verbs. The fact that they often occur before *yang* seems to suggest that they are still adjectives, and are a kind of predicative adjective.

The verb *mencengangkan* (stem = *cengang* "amazed") means "amazing", but more in the sense of being "wonderful" than causing someone to feel amazed.

Notes on Adjective Type 4 Verbs in this group are similar to Type 3, but the subject of the *-kan* verb is the *experiencer* and not the stimulus, unlike Type 3.

Notes on Adjective Type 5 Type 5 is like Type 4 except we never see Type 4 verbs in the pattern MEN+stem.

C.3 Noun Stems

A bare nominal can act as a predicate in *nominal sentences*, also referred to as *equational sentences*, which has the sentence structure: NP NP, as shown in Example (C.1).

(C.1) [from Sneddon (1996:233)]

Dia *guru.*
3sgteacher
"He is a teacher."

Unless in this kind of construction, nominals do not ordinarily predicate without derivation, which requires the affixing of a voice marker (see Section 2.2.3). Predicting the semantics of denominal verbs have been shown to be inherently idiosyncratic that the interpretation of denominalised are conventionalised and not entirely predictable (Jackendoff 2002:p35). This makes our task in producing *semantic scaffolding* around the stem to be even more difficult for nouns.

For this reason, we introduce 6 primes that can form a verbal unit with the nominal. These primes are PERFORM, ACHIEVE, MAKE, HAVE, CREATE, and BECOME. The predicate BECOME is part of the theory of Conceptual Structure, and is employed by Kroeger (2007) in describing the causative *-kan*. The other predicates were discovered as part of the process of mapping out and grouping syntactically-like verbs, and it is beyond the scope of this study to try to prove their cross-linguistic translatability to be incorporated into NSM.

With the addition of these 6 predicates, the noun types we arrive at are shown in Table C.3.

MORPHOLOGY	VERB FRAME	DECOMPOSITION
Type 1: *administrasi* "administration", *instalasi* "installation", *legalisasi* "legalisation", *nasionalisasi* "nationalisation", *ikat* "cord", *pukul* "blow/strike", *sewa* "hire".		
MEN+N_1	<NP_a, NP_b >	PERFORM-N_1([NP_a], [TO/ON NP_b])
MEN+N_1+KAN	<NP_a, NP_b >	PERFORM-N_1([NP_a], [TO/ON NP_b])
Type 2: *ajar* "lesson"		
MEN+N_2	<NP_a, NP_b >	HAVE-N_2([NP_a], [ON/TO NP_b])
	<NP_a >	HAVE-N_2([NP_a]) (unexpressed patient)
	<NP_a, VP>	HAVE-N_2([NP_a], TO VP)
	<NP_a, CP_{bahwa} >	HAVE-N_2([NP_a], THAT CP)
MEN+N_2+KAN	<NP_a, NP_b >	HAVE-N_2([NP_a], [ON/TO NP_b])
	<NP_a, PP_b >	HAVE-N_2([NP_a], PP_b])
	<NP_a, NP_b, NP_c >	PERFORM-N_2([NP_a], [TO NP_b ON NP_c])
	<NP_a, NP_b, VP_c >	PERFORM-N_2([NP_a], [TO NP_b TO VP_c])
	<NP_a, CP_{bahwa} >	HAVE-N_2([NP_a], CP)
Type 3: *gambar* "picture"		
MEN+N_3	<NP_a, NP_b >	MAKE-N_3([NP_a], [OF NP_b])
	<NP_a >	MAKE-N_3([NP_a]) (unexpressed theme)
MEN+N_3+KAN	<NP_a, NP_b >	MAKE-N_3([NP_a], [OF NP_b])
	<NP_a, PP_b >	MAKE-N_3([NP_a], PP_b])
	<NP_a, CP_{bahwa} >	HAVE-N_3([NP_a], CP)
Type 4: *aplikasi* "application", *ekspresi* "expression", *kerja* "activity/work"		
MEN+N_4	–	–
MEN+N_4+KAN	< NP_a, NP_b >	PERFORM-N_4([NP_a], [TO/ON NP_b])
Type 5: *belanja* "expenses", *gelembung* "bubble", *buku* "book", *publikasi* "publication", *radiasi* "radiation", *kumandang* "echo"		
MEN+N_5	–	–
MEN+N_5+KAN	< NP_a, NP_b >	CREATE-N_5([NP_a], [WITH/ON NP_b])

Table C.3: Noun Types 1-5

MORPHOLOGY	VERB FRAME	DECOMPOSITION
Type 6: *darat* "land", *didih* "boiling"		
MEN+N_6	$< NP_a >$	BE([NP_b, [AT N_6])
MEN+N_6+KAN	$< NP_a, NP_b >$	ACHIEVE-N_6([NP_a], [WITH NP_b])
Type 7: *hipotesis* "hypothesis", *titah* "command", *mimpi* "dream", *pikir* "idea", *tanya* "question"		
MEN+N_7	–	–
	<NP_a, NP_b >	MAKE-N_7([NP_a], [ON NP_b])
MEN+N_7+KAN	<NP_a, CP_{bahwa} >	MAKE-N_7([NP_a], CP)
	<NP_a, VP>	MAKE-N_7([NP_a], ON VP)
Type 8: *asumsi* "assumption", *umpama* "example", *wakil* "proxy", *lokasi* "location"		
MEN+N_8	–	–
MEN+N_8+KAN	$< NP_a, NP_b >$	MAKE-N_8([NP_a], [WITH NP_b])
Type 9: *paten* "patent" *tempat* "place" *tumpu* "foothold" *letak* "position" *penjara* "jail" *rumah* "house"		
MEN+N_9	–	–
MEN+N_9+KAN	<NP_a, NP_b >	CREATE-N_9([NP_a], [FOR NP_b])
Type 10: *injeksi* "injection", *kait* "hook", *analogi* "analogy" *maklumat* "declaration"		
MEN+N_{10}	–	–
MEN+N_{10}+KAN	<NP_a, NP_b, {PP_c} >	MAKE-N_{10}([NP_a], [OF NP_b]{ PP_c})
Type 11: *sesal* "regret", *susu* "milk"		
MEN+N_{11}	$< NP_a >$	HAVE-N_{11}([NP_a])
MEN+N_{11}+KAN	$< NP_a, NP_b >$	HAVE-N_{11}([NP_a], [FOR NP_b])
Type 12:*janji* "promise", *cerita* "news"		
MEN+N_{12}	–	–
	<NP_a, NP_b >	CREATE-N_{12}([NP_a], [ON NP_b])
MEN+N_{12}+KAN	<NP_a, VP>	CREATE-N_7([NP_a], TO VP)
	<NP_a, CP_{bahwa} >	CREATE-N_7([NP_a], THAT CP)
Type 13: *sarang* "web", *percik* "stain", *mula* "start", *kerja* "work"		
MEN+N_{13}	–	–
MEN+N_{13}+KAN	<NP_a, NP_b >	ACHIEVE-N_{11}([NP_a], [FOR NP_b])

Table C.4: Noun Types 6-13